目　录

本书所说的“统计学”，在我国习惯上常称之为“数理统计学”.对这些名词，存在着不尽相同的理解，这些将在本书的第 1 节中加以说明.

本书试图对统计学的学科性质、基本内容和发展历史，作一简要介绍.作为一本通俗性的书，又受篇幅和所使用的数学工具的限制，这个介绍在理论方面难于达到充分的严谨、全面和系统化，还请读者见谅.

本书前三节在数学上要求很少，一般具备高中程度数学知识的人，都可以读懂.第 4 节则要求读者具有一定的概率论知识，不过其中的基本思想，在第 1 节中已有所交代.最后一节，因系讲述历史，不能不提到某些在前几节中未充分解释的概念，不具备统计学初步知识的读者，浏览一过，大致有所了解就行了.

一、什么是统计学

什么是统计学？什么是数理统计学？这个问题，难于用简短的语言作一个明确、严谨而全面的回答.我们打算

先用通俗的语言作一个概括的论述，然后在适当的地方加以补充、解释，以使读者对此问题有一个比较清楚的理解.

实践是认识的来源. 所以,在研究一个问题时,往往首先要收集必须的资料.比方说,少年犯罪与家庭情况的关系如何?跑步对中、老年人的健康是否有益? 或更细致一些,对各种年龄的人,在什么时间,以怎样的速度,跑多长的距离为好?受教育时间的长短与其收入的关系如何?吸烟是否增加患肺癌的机会?增加多少?对一种产品的制造工艺作一些改变,是否有助于改善其质量，改善多少?凡此种种,都是很有现实意义的问题. 这类问题中,有的前人已作过一些研究,提出过种种见解.但前人的研究可能是在不同的条件下进行的(例如,对不同的国家，受教育时间长短与收入的关系就有不同),有的还受到当时科技发展水平和其他因素的局限(例如,某项研究由于经费的关系,收集资料的规模很小)，因此他们的结论不能照搬.如果所研究的是一个前人未接触过的新问题,那当然更不用说,收集资料这步工作是必不可少的.

收集资料的方法有两种:观察和试验.这两个词的含义的差别在于,在“观察”时,观察者可以说是处在被动的地位,他只是对所感兴趣的事物,记录下“自然而然地”发生的结果,而不去企图改变他所观察的事物.天文观察是一个典型的例子.在吸烟与患肺癌的关系的问题中,情况也是如此:你可以观察一个人是否吸烟,吸多少，观察他是否患肺癌;但你不能也不会去设法改变他的状况,这是

因为一个从不吸烟的人，不会应你的研究工作的要求而去吸烟．而在“试验”中，试验者则处在主动的地位，可在一定范围内自由地控制某些因素，以考察它们对其他因素的作用．典型的例子如在工业试验中，工艺参数如何取，原料配方如何选择，出自试验者的主动，以考察它们对产品产量和质量的影响．

从统计学的眼光看，观察和试验都是收集资料的方法．因此，许多统计学著作混用这两个词．但也应注意到，有些统计方法的合理使用和解释，与资料是来自观察还是试验有关．

在不少情况下，收集的资料可以用数量的形式表达．如一个人的月收入以人民币多少元计，是一个数字．有时需要研究事物的若干个方面，则资料可以用若干个数字（即一个向量）表达．如同时观察一个人的身高和体重，结果是一个二维向量．也有些情况，观察或试验所得只是事物所属的等级、类别等．例如观察一人的血型，结果为A、B、AB、O四类中之一；对一种酒品嘗结果，列入甲、乙、丙三等之一．这些，在必要时可以进行“数量化”．例如，约定把A、B、AB、O四种血型分别给以数字1、2、3、4．因此，在统计工作中，习惯上把所收集来的资料称为“数据”，或者用“样本”这个专门术语，意思都是一样的．

但是，认识并不是实践的直接产物．为研究一个问题而收集的资料，一般是一大堆杂乱无章的数字，从中看不出什么道理来．比如说，为研究吸烟与患肺癌的关系，观

察了 5,000 人，逐一记下每人日抽烟多少支，抽烟史多长，是否肺癌患者，患病多久等资料，订成一大厚册；泛泛翻阅这本册子，得不出多少东西．因此，需要把数据加以整理，从其中提取出与所研究的问题最有关的信息，并以简明醒目的方式表达出来．例如，一种可能的整理方式如下：把所观察的 5,000 人，按“不吸烟”、“每天吸 10 支以下”、“每天吸 10～20 支”、“每天吸 20 支以上”分组，从收集的资料中算出各组肺癌的发病率，并画成一张图；则很易看出肺癌发病率随吸烟量增加而上升的趋势，以及这个趋势的大小的概念．再举一个例子：为考察毕业后工作了 10 年至 30 年的大学生的工资收入状况，在这类人中抽取了 10,000 名进行观察，记下每人目前月工资数，得 x_1, x_2, $\cdots$, x_{10000} 等 10,000 个数据．计算其算术平均，即

$$\bar{x}=(x_1+\cdots+x_{10000})\div 10000,$$

就可以对这批人的收入的总的状况，或平均状况，有一个了解．如果要进一步了解收入参差不齐的情况，需引入另外的指标，例如

$$s^2=[(x_1-\bar{x})^2+\cdots+(x_{10000}-\bar{x})^2]\div 10000.$$

s^2 愈大，参差不齐的程度就愈大．这个指标 s^2 能给我们一些启示：s^2 太小，说明收入没有适当拉开档次，可能与平均主义有关；s^2 太大，则说明资历较浅的工作者收入太低，可能是反映了某种问题．因此，通过整理数据得到的这两个指标 $\bar{x}$ 和 s^2，以易于理解的方式告诉了我们不少东西(或者说，以简明的方式总结了这一大批资料的信

息）．如果想了解得更细致些，可以用一定间隔作单位（如1年，5年等），算出在这10,000人中，毕业年限在此间隔内的人，目前的月平均工资，用列表或画图的方式给出结果．自然，随着所研究的问题的不同、数据形式的不同，整理的方法也会有差别．可以说，在什么情况下该用怎样有效的方式去整理数据，属于统计学的研究范围．

按一定的方式整理数据，往往也就构成对数据的一种分析．例如，分析上例中的数据，可得出：毕业后工作年限每长一年，平均月工资增长多少．在吸烟与患肺癌关系的例中，分析所收集的数据，可以知道日吸烟支数每增加5支，或吸烟史每增加5年，肺癌的发病率增加多少．但无论对数据进行整理或分析，都没有越出所得数据的范围．就是说，分析所得的结果只对现有这批数据有效．就上例来说，设想分析结果是：大学毕业后每多工作一年，平均月工资多2元．这结果只是针对所抽查的这10,000人来说的．在我国，大学毕业后工作10～30年的，何止1万人．而我们真正感兴趣的，是这些工作人员的全体，而不止于所抽出作调查的这一小部分．这样，我们就需要向前跨一大步：基于所收集到的数据及对它进行整理分析的结果，对数据所来自的总体的有关情况，作出一定的论断．这种论断叫做“统计推断”．其具体形式，依问题中要求的不同而不同．所谓“数据所来自的总体”，就是指与所研究的问题有关的所有个体的集合．如在上例中，总体就是我国目前尚在的、大学毕业后工作了

10～30 年的全体工作人员．若这样的工作人员有二百万，则本问题的总体中包含二百万个体．被抽出作调查的那一万个体(即样本)，是其一部分．由于本问题中我们关心的不是具体的人而是其月工资，也可以说总体和样本分别由二百万个数和一万个数构成（其中可以有重复的）．这实际上就构成了一种抽象，且是很重要的抽象．因为这使我们可以摆脱总体及样本的具体属性，便于运用数学的方法，对不同的具体问题作出统一的处理方法．

如上文所述，统计推断的对象，是总体的有关情况，即因我们的研究目的而对之感兴趣的那些情况．在上例中，我们感兴趣的可以是总体平均值——即总体中二百万个工作人员的月平均工资 a，如果所抽查的那一万名工作人员的月平均工资为 $\bar{x}=90$(元)，那么未知的 a 是否就等于 90？当然不一定．但也易了解，这二者之间会有些关系．关系的大小和性质，取决于这一万名工作人员是如何抽得的．取决于从总体中所抽出的个体的数目(在此为 10,000)，它在统计学上称为“样本大小”——不难明白，样本大小愈大，$\bar{x}$ 与 a 一般就应愈接近．此外，还与总体的数学性质，即所谓数学模型有关．这一点留待后面再作解释．

统计推断的具体形式，依问题的要求而异．就此例而言，所要求的可能就是：根据样本，对总体平均值 a 作一估计．这种问题称为估计问题，是在理论上研究得最深入，在应用上最常见最重要的一类统计推断问题．总体平

均值 a 刻划了总体一个方面的性质，它称为总体的“参数”．因此，在统计学中，像估计总体平均 a 的这类问题，常称为“参数估计问题”．直观上觉得，应当用样本平均值 $\bar{x}$(在本例为 90)去估计 a．这个方法，即按样本算出的值去估计总体的相应值，是一个重要而常用的估计方法．

当我们声明，采用 $\bar{x}$ 去估计 a 时，我们就作出了一种统计推断．因为在这样做时，我们已越出了所掌握的样本的范围，而论及了样本所来自的总体．有的读者可能会问：这样一个看来似乎是纯粹形式上的步骤，能有多大的意义？其实不然．跨出这一步是不简单的．比方说，“用 $\bar{x}$ 估计 a”会有误差，这误差有多大，用怎样的形式表达出来，需要用到以深刻的数学理论为基础的统计方法．又如，用 $\bar{x}$ 估计 a 也并非理所当然的，唯一可行的方法．兹举一种可以设想的估计方法如下：把所得的 10,000 个数据按大小排序，取出居于正中的那两个，设为 x' 和 x''(若数据个数为奇数，则只有一个恰居正中，就以之代替下文的 x^*)，取其平均

$$x^* = \frac{1}{2}(x' + x'')$$

去估计 a．x^* 称为样本的“中位数”．用 x^* 估计 a，在直观上也是讲得通的．$\bar{x}$ 和 x^*这两个估计那个较好？在什么意义下较好？这是深刻的理论问题．除此而外，还可设想出其他种种在直观上看来也合理的估计方法．我们需要证明：$\bar{x}$ 这个估计在理论上具有某种优良性，这

样，用 $\bar{x}$ 估计 a 才有坚实的基础. 彻底解决这个问题，牵涉统计学上多方面的理论问题. 由此可见，正如我们所曾指出的，跨出这一步并非易事.

再举一例. 在研究吸烟与患肺癌的关系问题时，我们首先感兴趣的，可能是一个初步的问题：这二者到底是否有关，而暂不计其关系的深浅与其确切性质. 这问题可以较具体地解释如下：设如前所述，我们观察了 5,000 人，记录了各人是否吸烟与是否患肺癌的情况，经对数据作初步整理分析，觉得二者似有些关系. 但是，由于这 5,000 人只是地球上的成年人的很少一部分，仅凭这 5,000 人的数据而推及地球上全体成年人，有多大的可靠性？或更清楚地说，你在这 5,000 人中分析出的关系，是纯出于偶然性呢，还是确实反映了一种适用于全体成年人的规律性. 这个问题与上例中估计 a 的问题不同，它只要求回答一个"是"或"否"(是纯出于偶然，或否). 这种问题在统计学上称为"假设检验"问题. 它与参数估计，并列为统计学中两类基本推断问题，在理论上有深入发展且有重要应用. 名称的由来，是因为在统计学上处理这类问题时，先引进一个有关的假设. 如在本例中，引进"吸烟与患肺癌无关"这个假设. 然后，用样本去"检验"这假设是否成立. 具体说，我们通过分析所掌握的数据，看二者关系的大小如何：若关系不大，则不能排斥它是来自抽样的偶然性，因而断言二者有关的理由不充足，这导致我们接受上述假设；反之，若关系甚大，则仅以偶然性去解释是勉强的，因而有足够的理由断言二者有关，

这导致我们接受上述假设．这里，“关系大小”如何刻划，导致接受或否定假设的界限如何划分，都需要统计学的理论．除上述两种以外，还有许多形式更复杂的统计推断问题，需要以深刻理论为背景的不同处理方法．

由此可见，撇开收集数据的问题不谈，统计学的中心问题，或者说其主要内容，就是统计推断．统计学之所以有如此广泛的应用，正因为在数学上成功地发展了一整套有关的理论，并在其基础上，制定出了针对一些常见的重要问题的统计推断方法．就是收集数据的问题，也在一定程度上与统计推断的理论和方法有关．因为只有当数据的结构（这取决于用怎样的方式去收集数据）满足一定的条件时，才能对它运用适当的统计推断方法．不然的话，所收集的数据就不好处理．

以上在谈论统计推断问题时，我们是从一种科学研究的眼光去看待它，即它是以弄清事实为目的，不计较什么利害关系．有一类问题，通称为“统计决策问题”，或“统计判决问题”，与统计推断问题有关但又有差异．有关的地方是：统计决策问题的解决也要基于收集的数据，并使用统计推断理论中提供的种种方法．不同之处在于：决策（也常称为判决，或行动）要产生经济上的后果*)．在实际作出决策时，不仅要考虑到统计推断上的结果，还必须把经济上可能的后果结合进来．例如，有一批产品包含很多

---

*） 自然，决策的后果不必限于经济方面，但在统计决策理论中，只考虑那种问题，其后果可以用一定方法归结为经济上的得失．

件，要估计该批产品的废品率 p，则可以在该批产品中抽取若干个作检查，以样品中的废品率 $\hat{p}$ 去估计 p 即可．但是，如果这批产品是工厂对商店的供货，而商店经理要决定是否接收这批货，则问题并非简单地估计废品率 p 即可．因为，接收或拒收该批产品，都有经济上的后果．例如，若拒收，则当日无该货可出售，要损失利润；但如接收这批货，则有可能废品率 p 较大，而得不偿失．该经理作出的决策，除了考虑到 p 的估计值 $\hat{p}$ 以外，还须把每件废品的损失和出售每件合格品的利润结合考虑进来．另举一例：某工厂的设计试验部门，通过适当安排的试验并使用一定的统计推断方法，搞清楚了原料配方与产品性能之间的关系．但不同的配方涉及成本、原材料来源(这与运输费用也有关系)与消费者喜好，即市场前途问题，在最后作决策(即选用一种确定的配方用于生产)时，统计推断的结果自然是重要的．这只有在统计学家、专业人员、经济师和市场分析人员的共同参与下，才能作出适当的决策——当然，这中间涉及到的问题并非全是统计性的．

到这里，我们已说明了统计学是干什么事的．现在把它小结一下，而对统计学提出一个比较完整的定义：统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议．在这个定义中，有两点在上文未作仔细交代：一是“有效的方式”一语的含义，这涉及在收集数据的工作中具体的

作法问题．这个重要问题将在下文第二节作仔细论述．二是“带随机性的数据”一语的含义，对概率论初步知识略知一二的读者，自然明白其意义，下文我们还将略加解释．

此处引进的统计学定义，是依照《中国大百科全书·数学卷》中对“数理统计学”所下的定义，这个定义与《不列颠百科全书》上关于“统计学”的说法，基本精神也是一致的．后者把统计学定义为收集和分析数据的艺术．这个定义嫌过于简略一些．不过，其中“分析”一词兼有我们定义中整理、分析、推断的含义．它没有明确指出数据应带随机性，这是一个弱点(见下文)．至于此定义中称统计学是“艺术”，尽管有其不够严谨之处，却也有独到的地方：它提醒人们，统计学并不是一堆在应用时可以机械地照搬的公式，而是在应用上要发挥灵活性以至灵感，需要积累充分的经验．

按这个定义，统计学是一门与数字打交道的学科．在这个意义上，可以把它看成是数学的一个分支．它当然不是社会科学．还有一点要着重说明：像这样定义的统计学，在我国常称为“数理统计学”．而在西方，“统计学”和“数理统计学”有明确的区别，即数理统计学是统计学的数学理论那一部分．所以，在我国，数理统计学等于西方的统计学加数理统计学．其所以产生这个差别，与苏联对这个问题的看法有关．在苏联，把统计学定义为一门研究大量社会现象的社会科学，有很强的阶级性和党性；而数理统计学则被看成是在统计学中使用的数学方法及其理

论基础.这个看法对我国至今仍有很大的影响. 因此,在我国至今仍有不少人采取这样的看法:统计学是一门社会科学,数理统计学则是一门数学学科.

作者不打算在此对上述观点之间的分歧发表评论.然而,读者不难看出,本书是按照西方的观点来写的. 对此持异议的读者可以这样看待本书:它讨论了统计学与数学有关的那一部分.

按我们所讲的方式去理解统计学,自然地得出它的一个特点:它是通过事物的外在的数量上的表现,去揭示事物可能存在的规律性.它不能确认和解释,为什么事物会存在这样或那样的规律性,后者要依靠有关专门学科的研究. 不过,在探求这种规律性的解释的研究工作中,统计方法也有其作用.例如,用种种统计方法对一些统计资料进行分析的结果,都使人相信吸烟者中患肺癌的比率较高. 但是,究竟吸烟是引发肺癌的一个原因,还是这二者都受到同一遗传基因的控制?如果是后者,则统计资料分析的结果只是表明这二者有一种先天的联系,而不表明这二者有因果关系.要确定这种因果关系的存在,需要从医学上弄清吸烟引发肺癌的机制问题.

统计方法的这个特点,划清了统计学和其他学科的界线.例如,经济学、人口学、社会学、工程学、生物学……等学科,都用到统计学提供的方法.但统计学在这些学科中,只起着一个辅助性质的作用. 统计学自有其研究对象,即超脱了具体含义的数据的收集和分析问题. 当然,统计方法的这种辅助性质并不降低它的意义,恰恰相反,

由于事物的本质规律性往往隐藏很深，不易为人们所察觉，而其外在数量上的表现则易于引起人们的注意，以此，统计方法在揭示事物规律性的过程中，常能起到先导的作用.

按照上述观点，可以说统计方法是一种数学方法.在为数众多的数学方法中，统计方法有什么特点呢？因为，如果把统计学说成是一种处理数据的数学方法，那末，它与算术，一般讲与计算数学，就划不清界线．这里就要用到前面给统计学下定义时所加的那个限制词：随机性.统计学是处理带随机性的数据的问题．所谓随机性（又称偶然性)，是“随机会而定”的意思．从实际应用的角度去看，统计学中考虑的数据随机性有两种形式.一种形式的例子是前面提到的吸烟与肺癌关系问题，以及大学毕业后工作 10～30 年的人员的收入问题．在这些例子中，总体是由一些实在的个体(在此两例是人)组成．数据的随机性来源於，那些个体被抽出(以组成样本)，是随机会而定．举一个极端的例子．如果碰巧在你抽出的那 10,000 人中，大多数都是工龄短而工资高，或工龄长而工资低的人，则你会得出“工作年限愈长，收入愈少”的结论．虽则“碰巧”出现这类情况的机会不大，但既是抽查，你在逻辑上就不能绝对否定其可能性.由此也可以看到，统计推断有产生错误的可能.事实上，统计推断理论中的一个重要课题，就是计算在种种情况下，各种推断方法可靠的程度如何.

大体上说，这种随机性是与“观察”联系在一起的.另

一种形式的随机性则与“试验”相联系．简言之，就是常说的试验误差．例如，在一个天平上称一个物件，结果不会与物件的真实重量完全相同．误差的来源，除了一些较重大的、有可能指认的原因(如天平没有调准、制造上有缺陷之类)外，还有大量的无法指认和控制的偶然性因素．例如，邻近轻微的震动，操作者瞬间心情上的恍惚，等等．这使得重复称量得出的结果不尽相同．推而广之，在工、农业试验中，控制在一定条件下(一定的工艺参数，原料配方等)做试验，结果不尽相同且误差无从预料，这都表现为数据中的随机性误差．

“带随机性的误差”一语，在数学上有其确切的含义．确实，表面上看，数据中的误差，好像全然是杂乱无章的，看不出有任何规律性的东西．但是，我们要求这种数据在集体上显示出一种规律性，通过“概率论”中所谓“概率分布”来刻划．由概率分布所刻划的规律性，并不能规定或预言数据的值，而是大体上可以说成：它规定了这种数据产生的机制．举一个浅显的例子．在彩票开奖时，准备一个不透明的袋子，内装大小质地一样的 10 个球，上写有 0,1,…,9 等十个数字．彻底搅乱后，由一个蒙上眼睛的人抽出一个，登记结果后，放回去再彻底搅乱，再让他抽一个．这样可以一直继续下去，直到抽出需要的个数为止．在这个试验中，每步的结果都凭机会，无法预料．在这一点上说，无规律性可言．但试验的操作过程保证了，每个数字在每次抽取中，有同等的机会$\left(\frac{1}{10}\right)$被抽出．用概

率的语言说，每次抽取时，结果为任一指定数的概率，都是$\frac{1}{10}$.或者说，抽取的结果在 0,1,…,9 这些数字上呈均匀分布，这就是本试验中，数据的概率规律性.

确切地说，统计学理论和方法，是建立在数据具有这种概率规律性的假定的基础之上的. 不满足这一点的数据，无法用统计方法去处理. 但在每一具体问题中，即使我们有理由假设这种规律性存在，往往也不易于确定其形式如何. 这要求有所研究问题的专业知识、经验，有时则多少是一种数学上的假定.不过，统计学上也发展了一些方法，帮助我们根据数据提供的信息去确定这种规律的形式，即数学模型，或者帮助我们验证，某种在理论上假定的数学模型是否与实际(即数据)相符. 另外，也有些统计方法有较广的使用范围，而不甚依赖数学模型的确切形式.

由以上论述，不难看到统计学与概率论的密切关系. 尽管也有些统计学著作，不依赖或基本上不依赖概率论的概念和方法，但这些著作只能介绍统计学的一些方法，就是说，告诉你怎样去做. 一涉及这些方法的道理，还得乞灵于概率论. 总之，现时还没有找到一种既摆脱概率论，又能严密完整地阐述统计理论的方法. 以此之故，人们常说概率论是统计学的基础，统计学是概率论的应用. 这个说法正确地概括了二者关系的基本方面. 不过应当明确，统计学与概率论是两个平行的姊妹学科，并无高低从属之分. 正如解析几何学中大量使用了代数方法，但由

于有其自身的问题与特点，它被公认为一门不从属于代数学的学科.

关于统计方法与其他数学方法的界线划分问题，上面的论述集中在“随机性”的有无这一点上(应当说明的是:在有些数学分支,如“运筹学”中，也常讨论涉及随机性的问题，这应理解为概率统计方法在这些学科中的应用,而并非这些学科的根本性特点). 现举一个更实际的例子来说明. 设有一块正圆柱形的均匀木头，其比重 a 和高 b(厘米)都假定为已知(不带误差)，现想要知道其半径 r，但身边没有尺子,只有一把秤,称一下这木头,得其重量 A(克),于是由公式 $A=ab\pi r^2$ 算出半径

$$r=\sqrt{\frac{A}{ab\pi}}\text{(厘米)}.$$

若秤毫无误差,则整个解题过程,不过是一个几何公式的应用,与统计学无关.反之,若称量结果有随机性误差(一般当然都如此)，而我们在精度要求上又较高,则需要把这木头重复称若干次,得结果 $A_1,\cdots,A_n$. 以样本平均值

$$\bar{A}=\sum_{i=1}^{m}\frac{A_i}{n}$$

作为木头真实重量 A 的估计值. 然后,用

$$\bar{r}=\sqrt{\frac{\bar{A}}{ab\pi}}$$

作为半径 r 的估计值. 在这里，统计方法就参与了解题过程. 例如，用 $\bar{A}$ 估计A，是统计上惯用的方法. 尤其是,用 $\bar{r}$ 估计 r 的误差如何,需通过用 $\bar{A}$ 估计 A 的误差及关系式

$$\bar{r} = \sqrt{\frac{\bar{A}}{ab\pi}}$$

去考察．这也涉及概率论和统计学的方法．

在结束这一节之前，还想在统计学与数学的关系这个问题上补充几句．统计学与概率论之密切关系已如上述；概率论是数学的一个分支，这一点，数学各分支(包括概率论)的学者都无异议；至于统计学是否应称作是数学的分支，情况就较复杂一些，我们撇开那种认为“统计学是一门社会科学”的意见不谈(承认这一点，自然就不能认为它是数学的一个分支)，现结合作者本人对这个问题的想法提供几点意见，供读者参考．

1．如果把数理统计学理解为其狭义解释，即它是统计方法的数学理论基础部分，则它可视为数学的一个分支．这一点，在中外统计学家，包括那些认为统计学是社会科学的学者中，似乎是没有分歧的．

2．如果把数理统计学按我国一般习惯上所作的那种广义解释，即西方意义下的统计学(以收集和分析数据为任务)，则既然所涉及的数据已超脱了具体含义，把这种较广意义下的数理统计学理解为数学的一分支，看来也还是恰当的．的确，现在有一派观点，主张不拘泥于数学模型，而主张依靠电子计算机等先进工具去处理数据，提取有关信息，以探求适用面更广泛的方法．但如这一切需要上升为理论和系统化，恐怕终究不能不借助于数学理论．如说它自成一实体，则其确切性质如何，需要说明．

3. 可是，在作为现代统计学发源地的英国(约自本世纪初始)，以及在目前统计学最发达的美国(其统计学大规模发展大致始于本世纪三十年代)，统计学一直是在不从属于数学的情况下发展起来的.他们在很早的时候，就在大学里建立了与数学系并列的统计系，成立了专门的学会与研究所，出版了多种统计学的专门杂志. 这些表明他们是把统计学看成与数学并列的学科. 造成这种情况的原因不在此细论，不过，这更多的是与怎样能使统计学得到更好的发展有关，还不能看作是西方统计学界对“统计学与数学的关系的性质如何”这个问题的回答. 实际上，在西方统计学家的著作中，不大直接涉及这个问题.

在我国，统计学究竟是数学的一个分支，还是与数学并列的一个学科的问题，也在逐渐引起人们的关心.不时听到有关这问题的种种议论.然而，尽管这无疑在理论上是一个重要而有趣的问题，但更现实的问题是：统计学的发展以采取怎样的组织形式为好. 这里面有些是随着学科的自然发展而解决的，有些则涉及某些政策上的问题. 这话离题太远，不属本书范围.

二、抽样调查与试验设计

在第一节已说过，统计学的任务是有效地收集资料

(即数据),并对之进行处理(整理、分析、推断等). 本节将对前一任务作较仔细的讨论.

我们曾指出,收集资料有两种方式:观察和试验. 与此相应,在统计学中产生了两个分支学科, 一曰"抽样调查"或"抽样技术", 一曰"试验设计". 因此, 我们也就分这两个专题来介绍.

1. 抽样调查

前面提到的关于大学毕业后工作 10～30 年的工作人员收入状况问题,可用来解释一般的概念.先把有关之点列举出来:

1. 我们的研究工作所关心的个体,有一个明确的范围.每一个体是一个"看得见,摸得着"的实体.

2. 所有这样的个体组成一个总体. 这总体所含个体,尽管为数甚大,但仍是有限的.

3. 我们所关心的,其实并不在于每一个体本身, 而在于它的某些指标值.

如在此处, 关心的是两个数字: 毕业后工作了多少年,目前月工资如何.

4. 研究的目的是弄清该总体的某种性质.比如, 这批人平均工资如何,工资散布的程度(不齐性)如何,或一般地说,各种年限内部工资分散的程度如何,不同年限间分散程度如何等等.要明确的是:尽管这些性质离不开每一个体的指标值, 但它是有关这全体指标值的一种集体性质,不从属于任一特定的个体.

5. 由于总体中所含个体数太多,无法对之一一作调查,而只能抽取其一部分作调查,以其结果去推断上面第4条中提到的那些性质.

最后这一条,是我们目前讨论的对象. 但应当了解:其余几条也不总是不言自明地容易处理的.例如,考察怎样的指标与我们研究的问题最贴切?就本例来说,是只虑其正式工资收入为好,还是把其他收入也考虑考进来为好?表面上看,后者似乎较为合理. 但是,各人目前非工资收入该如何定,能否比较确切地调查出来,还有,有些人的某种收入(如既是职工又经商)与其学历和工龄似乎无关,是否也该计入,等等,这些都是问题. 且这样一来,工作量势必大大加重. 其次是调查的技术. 如所调查的指标涉及被调查者私人的状况或看法,你如何使他无顾虑地把真实情况提供出来. 在理论上,我们假定的模型是:每一个体有一个或多个已知而确定的指标值与之相联,但在实际中决没有这么简单. 第三,总体范围该定得多大. 太小则研究结论适用面很窄,太大则有人力、物力等问题. 以上这些问题的妥善考虑与处理,都关系到研究工作的成败或成功程度.

在作了这些交代后,我们来讨论上述第5条中提出的抽样问题.这方面的材料虽被写成整本的大著作,但从原则上看说来却很简单:统计学根据人类在这方面长期积累的经验,配合概率论理论上的要求,提出了一条基本要求:抽样要保证对每一个体"机会均等",即总体中每一个体有同样的机会被抽到,谁也不占优先.凡是适合这个

原则的抽样，在统计学上叫做"随机抽样".我们设想一种实现这种抽样的具体作法如下：设总体中共有 N 个个体，需要抽出 n 个，把这 N 个个体分别编号为 1, 2, …, N.准备 N 个大小、质地一样的球，分别在其上书写数字 1, 2, …, N，将它们放在一个不透明的口袋中，彻底搅乱后，从中一次抽出 n 个，或一次抽一个，抽出的不再放回，直到抽满 n 个为止.凡是其编号在抽出的那些球上的个体，组成我们的样本. 在这种抽法下，每一个体只能在样本中出现一次，因而称为"无放回的抽样". 也容易看到：从 N 个个体中(不放回地)抽 n 个，不同的结果有

$$C_N^n=\frac{N!}{n!(N-n)!}$$

种，它们都具有同等的出现机会. 这比只要求每一个体有同等机会被抽到要更进一步. 为表达这些性质，在统计学上有时把这种抽样叫做"无放回简单随机抽样".

在实际应用中，抽样几乎都是无放回的，即要求同一个体不能在样本中重复出现. 若因为某种原因需要破除这一限制，则必须一个一个地抽，每次抽出球后，登记其上的数字，放回袋中，彻底搅乱再抽下一个，直到抽出 n 个为止. 这种抽样叫"有放回的"，其样本大小 n 可以超过总体所含个体数 N. 在无放回抽样时，当然有 $n\leqslant N$. 有放回的抽样在理论上比无放回抽样简单，且在比值 n/N 甚小(例如，不超过 0.05)时，两种抽样方式的差别，从实际观点看并不重要. 因此，有放回抽样在抽样理论中占有一席地位.

如果总体中所含个体数很大，则按上述“口袋模型”去操作，很不方便．为克服这个困难，人们设计了一种叫“随机数表”的东西，来代替这个口袋．不妨设想：在一个口袋中放了10个球，其上分别书写0,1,…,9这10个数字，然后有放回地一个一个地抽球（每次抽后彻底搅乱），并将其上的数字依次排列在一本书的各页上，就成为一本随机数表．比如下表是一本随机数表之一页的一部分：

05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 58 78 06 83
09 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93
68 71 86 85 85	54 73 32 08 11	12 44 95 92 93	16 29 56 24 29
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74
68 65 52 14 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50
17 53 77 58 71	71 41 61 50 72	12 41 94 96 26	44 95 27 36 99
90 26 59 21 19	23 52 28 33 12	96 93 02 18 39	07 02 18 39 10
41 23 52 55 99	31 04 49 69 96	10 47 48 45 88	47 48 45 88 85
60 20 50 81 69	31 99 73 68 68	35 81 33 03 76	41 53 03 96 81
91 25 38 05 90	94 58 28 41 36	45 37 59 03 09	90 35 57 29 12
34 50 57 74 37	98 80 33 00 91	09 77 93 19 72	74 94 80 04 04
85 22 04 39 43	73 81 53 94 79	33 62 46 85 28	08 31 54 46 31
09 79 13 77 48	73 82 57 22 21	05 03 27 24 83	72 89 44 05 60
88 75 80 18 14	22 95 75 43 49	39 82 82 12 49	02 48 07 70 37
90 96 23 70 00	39 00 03 06 80	55 85 78 81 36	94 37 80 69 52

设想总体中包含70个个体，要无放回地抽出10个．把这70个体自1至70编号后，随意翻开随机数表，用手任意在上面一指，设指到“13”，则用该表的第13页，设即为此处列出的那一页，把两列合并自上至下，自左至右读去，依次得05,09,68,26,68,17,90,…．其中68出现两次，只取其中一个(因不放回)；90因大于70，也不要．

按这个方式，得到 05,09,68,26,17,41,60，34，65，53．即抽出了以这些数为编号的个体．若总体中所含个体数在 100 到 999 之间，则需合并三列，余类推．当然，使用随机数表并不能免除将总体中的个体加以编号这个麻烦．

对有些人来说，总感到随机抽样这种思想难于接受．他们认为，既然所要达到的目标，是使抽出的那些个体(样本)能够尽可能地代表整个总体的情况(这是正确的)，那末，通过有计划的、自觉的安排．而不委之于随机性，岂不能更好地实现这一点．问题在于，这种作法难于免除一般人都多少会有一些的主观偏见，特别是在研究者希望得到某种结论时，更是如此．另外，除非总体所含个体数 N 很小(这时，抽样调查大概没有必要)，人们很难掌握总体中的个体的有关情况，而人为的安排可能导致重大偏差．反之，根据概率论中的大数定律，随机抽样的方式，保证了当样本大小 n 较大时，总体中具有各种性质的成分，各按其比率均衡地出现在样本中，因而在这个无形的“自然调节”中实现了所企求的代表性．实际应用的经验也证明了这一点．

工作中的图方便，以及考虑不周，往往是破坏抽样的随机性的重要原因．如派某甲去一个县调查农民收入情况，他为图方便，只在县里交通较便利的河流、公路沿线挑选若干户作了调查．由于交通方便的地方，农民收入一般也较高，某甲的抽样调查结果将不能反映全县农民的真实情况．下面是一个著名的例子：1936 年美国大选，由

民主党人罗斯福对共和党人兰登．美国有一家著名杂志作了大规模的民意测验，共调查千万人以上，作出回答的二百余万人，据其结果，该杂志预言兰登将以压倒优势获胜．事实上，结果完全相反，罗斯福以压倒优势胜兰登．原因在于，该杂志是从电话号码簿和俱乐部名册等去选择被调查者，这类人多属于富有阶层，倾向共和党者多．另外，大量的“无反应”情况(约八百万人)也造成了显著的偏差．这后一点，在抽样调查工作中是值得注意的．

根据情况的需要，在实际运用时，上述简单随机抽样方案有时要作些变通．重要的有以下两种：

一是“集团抽样”．即先把总体中的全部个体，按某种考虑分成一些大集团．每个大集团内又可分为若干个小集团，后者还可以再细分．抽样时，先用随机化的方法抽取若干个大集团，再在抽出的每个大集团内，分别抽出若干个小集团……这样下去，最后在最低一级的集团中，随机抽出若干个体．这样抽出的全部个体，构成我们的样本．这种作法，是为了防止样本中的个体在地域上过于分散，而过分地加大了工作量．举例言之，设要通过抽样调查去了解某县农民收入情况，该县农户以十万计，计划从其中抽出 400 户．若按简单随机抽样的方式去抽，则这 400 户可能散布在全县每一角落，逐一访问至为不便．为缓和这一点，可改用如下的抽法：先在全县随机抽出若干个乡；在抽出的每个乡中，各随机抽出若干个村；最后，在抽出的每个村中，各随机抽取若干农户．这样，最后抽出

的农户相对集中一些，而又不甚影响随机性.在涉及到全国规模的抽样调查中，这种作法更不可免.

二是"分层，按比例". 举例言之，设要了解目前大学教师的收入情况，在全国，这种人以十万计，而我们只能调查其一小部分，例如 1000 人.在使用随机抽样方法时，由于抽取的量(1000)不算很大，大数定律的均衡作用未能充分发挥，于是在样本中，可能会出现各阶层人员的比例与总体中的比例有相当偏离的情况，而这就会影响样本的代表性. 例如，若在样本中老教授偏多，则调查结果将偏高.为补救这一点，设计了"分层，按比例"的抽法：把大学教师按现行职称序列分别助教、讲师、副教授、教授四层. 设已知这四层的人数比例大概是 15%、50%、30%、5%(这是随意假设的数字). 预定抽出 1000 人，按上述比例，各层应抽出人数为助教 $1000\times15\%=150$ 人，讲师、副教授、教授分别 500 人、300 人和 50 人.然后，在各层内，用简单随机抽样的方法，抽出所需的人数. 抽出的属于各层的人即组成样本.

在这一抽样方案中，既有计划的部分，又有随机会而定的部分.计划部分(分层，按比例)对机会的影响和作用作了"宏观"上的控制，而在"微观"(各层内)上，则让机会起调节作用.

这种作法的目的，不言而喻，是为了限制机会的破坏作用，以使样本达到更好的代表性. 必须指出，这与前面批评过的那种按主观指定样本的作法，毫无共同之处.这里的分层，是有客观依据的，并非由人们主观上觉得如何

而定.不过,为保证这方法有效,有两个条件:一是分层的标准应合理. 比方说,在此例中,若不按职称分层而按学科分层,或按省分层,则因各学科或各省大学教师的工资差别不大,这种分层就无益. 二是每层所含个体数在总体全部个体数中所占比例必须比较确切地知道. 若不然,而由人主观想象定一个比例,则反而会引进系统性的偏差. 例如,错误地把"教授"这一层的比例定为50%,而在样本1000人中包含进500名教授,结果会系统地偏高.

这个例子的思想,直接推广到一般情况:分层的标准是,使每层内各个体指标值变化尽量小,而不同层之间,个体指标值的变化尽量大.能实现这一点,就是成功的分层方法. 另外,分层法与集团抽样可以联合使用:每层内可以分集团,集团内也可以分层. 用这种方式,就可以构造出种种复杂的抽样方案. 当然,抽样方案的选定,要考虑实际问题的条件和需要.

2. 试验设计

一般的提法如下:有一个(或几个)我们感兴趣的指标,如工、农业产品的质量或数量;以及若干个我们选定的,对此指标可能有影响的因素或变量,试验的目的是考察这些因素与指标的关系.如某一因素对指标有无影响,影响多大,各因素处在何种状况下对指标值最有利,等等. 例如,在种植玉米时,有四个品种和三种肥料可供选用. 则这试验中有两个因素:一是种子品种,它有4个不

同的状态，每个状态(即一种具体的品种)称为品种这因素的一个水平，故品种这因素有 4 个水平(称为 4 水平因素)；一是肥料，它有三个水平．本试验的指标值可定为亩产量(斤数)．试验的目的是弄清不同种子品种与肥料对产量有无影响，多大影响，如何选择这两因素的各一个水平，使产量最高，等等．

（一）单因素试验

只包含一个因素的试验，称为单因素试验；否则，称为多因素试验(具体有二因素试验，三因素试验等)．下面先讨论单因素试验的设计问题．

先由种种考虑(所要求的精度，人力物力条件等)定下总的试验次数 n．把这 n 次试验分配给因素的各水平，在可能的条件下，总是平均分配．如 $n=15$，因素有 3 水平，则各水平做 5 次．为此，要准备 15 份试验材料(或称试验单元)．一般，设有 c 个水平(分别编号为 1，2，…，c)，各水平分别预定作 n_1，n_2，…，n_c 次，有 $n_1+n_2+\cdots+n_c=n$．

设计的问题，集中到一点，就是如何把这 n 份试验单元分配(按 n_1，n_2，…，n_c 的数目)给这 c 个水平．例如，3 个玉米品种，准备了15块试验田，每品种 5 块，具体如何分法．这要看试验单元的情况而定．由于这一点，产生了种种的设计方案．今介绍几种如下：

1．完全随机化设计．即纯粹凭机会去分配试验单元．具体作法如下：先把 n 个试验单元分别编号为 1，2，…，c．然后用随机数表，从其中无放回地抽出 n_1 个给水

平1，再在剩下的 $n-n_1$ 个中抽 n_2 个给水平2，等等. 具体说来，如设 $n=15$，$c=3$，$n_1=n_2=n_3=5$. 使用前面列出的那页随机数表，并两列由上至下、由左至右读，大于15（以及00）不要，重复的不要，则先读出5，9，10，4，13. 这几号试验单元给水平1. 再往下读，挑出14，3，15，8，6这几号试验单元给水平2. 余下的给水平3.

这种设计适合于各试验单元条件比较均匀的情况. 若不然，则在可能的情况下，应采取前面介绍过的分层方法，见下文.

2. 完全随机区组设计. 先举一例. 设上述玉米种植试验在五个村子里进行，每个村子提供3块面积形状一样的试验田. 但同一村的三块地条件较均匀，而不同村的地块条件差别较大. 这时，若用完全随机化设计，则某些品种可能碰巧都分给条件较差的村子里，而不利于该品种. 为避免这一点，我们把每村子里那三个试验单元作为一"层"，而规定在每一"层"内，三个品种必须各占一块地，至于那一个占那一块，则由随机的方式决定. 这一设计安排，就免除了上述可能性.

我们把一层内的三个试验单元，称为一个"区组". 一般地，把条件接近的一组试验单元称为一个区组. 若因素有 k 个水平，则每个区组必须包含 k 个试验单元，而全部 n 个试验单元应能分解为 r 个区组：$n=kr$. 每个水平在每一区组内恰占一试验单元；具体占那一个，则纯由随机化确定. 这种试验安排，就叫做"完全随机化区组设计". "完全"是指每区组都包含 k 个试验单元，而 k 即为

因素的水平数.

这种设计在实用上很常用.这是因为,在规模较大的试验中,要弄到足够多的均匀的试验材料,不是易事."区组"一词在实际应用中有很广的意义.再举一例.为比较4种原料配方的优劣,对每种配方,各准备了 5 份材料作试验;而参与试验的有 5 人,其操作水平高低不一.为避免因此而造成的误差,可让同一配方的 5 份材料中,每人各操作一份,具体分配则由随机化确定. 在此,可以说参与试验的每个人都构成一个区组.

3. 平衡不完全随机区组设计. 有时,区组所含试验单元数 t 小于因素的水平数 k.这时,无法在每一区组内把因素的各水平都做一次试验.这种区组称为"不完全区组".在前述玉米种植试验中, 若每个村子里只给两块试验地,就有一个不完全区组的设计问题.这种设计所追求的,是在区组不完全的困难条件下,设法达到某种程度的平衡.兹举一例说明之.有 5 个玉米品种,在 10 个村子里进行试验,每个村子提供 3 块大小形状一样的试验田.这时, $k=5$, $t=3$.考虑如下图的设计安排:

1 2 3	5 1 3	4 5 1	2 3 5	3 4 5
1	2	3	4	5
3 4 1	4 5 2	2 3 4	1 2 4	5 1 2
6	7	8	9	10

每一个框框内的三个数字,表示同属一个村那三块地(即一个区组)所种植的三个品种. 那个"框框"分配给那个村, 每个村内那三块地如何分配, 都按随机化的方式决定.细察这个设计,有以下几个特点:

① 每区组中都含 3 个不同的水平；

② 每个水平都在 6 个区组内出现；

③ 任一对水平在同一区组内同时出现的次数都是 3.

例如，水平 1、2 同在区组 1、9、10 中出现，水平 3、5 同在区组 2、4、5 内出现，等等. 这几个性质标志了一种平衡的特点. 又因其区组为不完全，且在区组内实行随机化，而得出本设计的名称——平衡不完全区组设计，简称 BIB 设计. 如上所述，一个 BIB 设计有 5 个参数：

k——因素水平数；

t——每区组所含试验单元数，常称为区组大小；

b——区组数，在此例为10；

r——每个水平的试验次数，在此例为6；

λ——任一对水平在同一区组内同时出现的次数，在此例为3.

这 5 个参数要满足一些条件：

$$bt = kr, \quad \lambda(k-1) = r(t-1), \quad b \geqslant k.$$

前两个等式很易证明；后一不等式是试验设计的奠基者费歇耳(R. A. Fisher)得到的，证明比较困难. 而且，这三个关系也不是 BIB 设计的存在的充分条件. 直到现在，这个问题仍未完全解决.

4. 拉丁方设计. 先说明什么叫拉丁方. 考察下图中由数字 1,2,3 构成的三阶方阵：

$$\begin{matrix} 2 & 1 & 3 \\ 3 & 2 & 1 \\ 1 & 3 & 2 \end{matrix}$$

发现它有这样的特点：在每一行及每一列内，数字 1,2,3 各出现一次．因这个性质，称这个方阵为“三阶拉丁方”．对任何自然数 n，不难构造出 n 阶拉丁方．例如，第一行依次写 $1, 2, 3, \cdots, n-1, n$．第二行自 2 开始，余类推，如图所示，即得一个 n 阶拉丁方．

$$\begin{array}{cccccccc}
1 & 2 & 3 & \cdot & \cdot & \cdot & n-1 & n \\
2 & 3 & 4 & \cdot & \cdot & \cdot & n & 1 \\
3 & 4 & 5 & \cdot & \cdot & \cdot & 1 & 2 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
n-1 & n & 1 & \cdot & \cdot & \cdot & n-3 & n-2 \\
n & 1 & 2 & \cdot & \cdot & \cdot & n-2 & n-1
\end{array}$$

当 n 较大时，不同的 n 阶拉丁方为数很大．现在尚未弄清楚这个数的确切公式．

拉丁方用在田间试验中，起着所谓“双向区组”的作用．例如，有三个玉米品种，在一块长方形的试验田上进行试验，将其分为 9 等分，每个品种占 3 块．若这块地的肥沃程度和其他条件沿两个方向都有差异，则按三阶拉丁方设计如图：

<table>
<tr><td>2</td><td>1</td><td>3</td></tr>
<tr><td>3</td><td>2</td><td>1</td></tr>
<tr><td>1</td><td>3</td><td>2</td></tr>
</table>

则任一品种在任一方向上都不占优势．在工业试验上拉丁方设计也有用，见后．

上面讲述的各种设计，包含了三个要点：一是分区组以在“宏观”上控制系统误差；二是在区组内实行随机化，以在“微观”上避免主观因素引起的误差；三是实行重复

(即每一水平做若干次试验)以缩小试验误差的影响. 这就是费歇耳提出的试验设计三大原则.不言而喻,这些原则的精神也可用于多因素试验*). 现在我们就转向讨论这种试验.

(二) 多因素试验

若试验中包含 N 个因素，分别具有 k_1、k_2、…，k_N 个水平,则这一试验称为一个 $k_1\times k_2\times\cdots\times k_N$ 试验. 若 $k_1=\cdots=k_N=k$,则简称为 k^N 试验. 试验时,把每一因素各取一水平,组成一个“处理”，将其施加在试验材料上.如前面提到过的那个玉米种植试验，有 4 个玉米品种，3 种肥料,这是一个 4×3 试验,共有 12 个处理. 每一处理由一个选定的品种和肥料组成，即在一块试验地上种植该品种并施放该肥料.

一个 $k_1\times\cdots\times k_N$ 试验包含 $t=k_1\cdot k_2\cdots k_N$ 个处理.若每个处理都做一次试验,则称为本试验的一个“全面实施”. 除非 N 和 k_1，…,k_N 都比较小,则 t 将相当大. 因此,在多数情况下，全面实施是不现实的，而只能取 t 个处理中的一部分去做,称为“部分实施”.一般都是取一自然数 d(能整除 t),而取 t/d 个处理做试验，称为“$1/d$ 部分实施”.部分实施的困难之处，在于要使得所选出的那一部分处理，在各因素各水平间保持一定的平衡**). 为明白这一点,举一简单例子,设有 A、B、C 三因素，各 2

---

*) 一般,多因素试验需要做的“处理”(意义见下文)较大,故不常实行重复.

**) 确切含义见下文.

水平．这是一个 2^3 试验．设只作其 1/2 实施．把每一处理写成(ijk)的形式，表示 A、B、C 的水平分别取 i、j、k．考虑两个部分实施方案甲、乙：

$$甲：\underset{x_1}{(111)},\underset{x_2}{(112)},\underset{x_3}{(121)},\underset{x_4}{(212)}.$$

$$乙：\underset{y_1}{(111)},\underset{y_2}{(212)},\underset{y_3}{(221)},\underset{y_4}{(122)}.$$

每个处理下的量，表示该处理的试验结果．例如，x_1 表示在方案甲中，处理(111)的试验结果．余类推．

设用方案甲，而需要比较因素 A 的两水平 1、2 的优劣．水平 2 只有一个试验结果，即 x_4；水平 1 虽有三个试验结果，但可与 x_4 相比者，唯有 x_2，因为在其余两个试验结果中，所涉及的因素 B、C 的水平与 x_4 的不一样．这样，虽做了 4 次试验，但我们只能用上两个．

若用方案乙，则这个比较可通过$\frac{1}{2}(y_2+y_3)-\frac{1}{2}(y_1+y_4)$去进行．因为，比方说，在$\frac{1}{2}(y_1+y_4)$中，因子 B、C 的水平 1、2 各出现 1 次，在$\frac{1}{2}(y_2+y_3)$中也如此．故差$\frac{1}{2}(y_2+y_3)-\frac{1}{2}(y_1+y_4)$只反映了因素 A 的 1、2 水平的差别，B、C 的作用完全抵消了．在这里，4 次试验结果都用上了，其所以能做到这一点，是因为方案乙是根据一定的方法选出，保持了各因素各水平之间的平衡．

另外，在多因素试验中，也有划分区组的问题．这种

划分，同样也要保持各因素各水平之间的平衡．不然的话，区组间的差别就会与因素的效应混杂起来．不过，从原则上说，划分区组的问题也可看成一种部分实施的问题．因为，只须把“区组”本身作为一个因素，其水平数即为区组的个数．

那末，用怎样的方法可以实现这种有平衡性质的部分实施设计呢？这里介绍两种常用的方法．

1．拉丁方和正交拉丁方．

拉丁方可用于 n^3 型试验的 $1/n$ 实施．这包括 n^3 型试验的全面实施，但分 n 个区组，每区组包含 n 个试验单元．为确定计，举 $n=4$ 为例．选定一个 4 阶拉丁方，如下：

$$
\begin{matrix}
1 & 2 & 3 & 4 \\
2 & 1 & 4 & 3 \\
3 & 4 & 1 & 2 \\
4 & 3 & 2 & 1
\end{matrix}
\tag{1}
$$

设有 3 个因素 A,B,C，都是 4 水平．对这拉丁方中的每个元素，写出行号，列号，该位置的数字．例如，第 2 行第 3 列处数字为 4，故上述三元组为(234)．依此方法，自第一行始，全部 16 个三元组为

$$
\underset{1}{(111)},\ \underset{2}{(122)},\ \underset{3}{(133)},\ \underset{4}{(144)},\ \underset{5}{(212)},
$$

$$
\underset{6}{(221)},\ \underset{7}{(234)},\ \underset{8}{(243)},\ \underset{9}{(313)},\ \underset{10}{(324)},
$$

$$
\underset{11}{(331)},\ \underset{12}{(342)},\ \underset{13}{(414)},\ \underset{14}{(423)},\ \underset{15}{(432)},\ \underset{16}{(441)}.
$$

这就是我们所定的、包含全部处理数 $4^3=64$ 的四分之一

的那个部分实施．仔细检查一下，会发现它有这样的性质：在任一因素的任一指定水平的那些处理中，其余各因素的各水平都出现一次且只出现一次．例如，因素 B 的水平 2 在 2、6、10、14 等几号处理中出现．在其中，因素 A 的水平 1、2、3、4 分别在 2、6、10、14 号处理中，因素 C 的水平 1、2、3、4 分别在 6、2、14、10 号处理中．这就是我们前面多次提到的“平衡”性的确切含义．由于设计有了这种平衡性质，当任一因素的任两个水平进行比较时，可以把包含这两个水平的全部试验结果都用上，且不受其他因素的干扰．这一点，在前面曾作过解释．

如果只有两个 4 水平因素 A、B，但要分 4 个区组做，则只须把刚才的因素 C 看作区组，它的同一水平的处理列入一个区组内．这样，由上述设计，得到 4 个区组的划分为：

区组 1：(11)，(22)，(33)，(44)；

区组 2：(12)，(21)，(34)，(43)；

区组 3：(13)，(24)，(31)，(42)；

区组 4：(14)，(23)，(32)，(41)．

它具有如下的特点：每个区组包含 A、B 的 4 个水平各 1 次．因此，任一因素的任一水平，都不致因区组的划分而处在有利或不利的地位．

若因素个数大于 3，或等于 3 而要分区组，则一个拉丁方已不够，必须使用几个具有所谓“正交”关系的拉丁方．例如，前面的 4 阶拉丁方 (1) 与下面列出的拉丁方 (2)：

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

(2)

构成一对 4 阶正交拉丁方．“正交”的意义为：在(1)中同一数字所占的 4 个位置，在(2)中则恰好各数字都出现 1 次．现设有 4 个 4 水平因子 A、B、C、D，要作其 $1/4^2$ 即 1/16 部分实施．为此，只须就方中的每个位置写出 4 元组(行号，列号，(1)中的数，(2)中的数)．例如，对第 2 行第 4 列这位置，此 4 元组为(2432)．这样写下的 16 个处理，就构成所要的部分实施．如果只有三个 4 水平因素 A、B、C，但要分 4 个区组，则只须把因素 D 的各水平作为区组号即可．若有 5 个 4 水平因素，而要作 $1/4^3=1/64$ 实施，则要用到三个互相正交的拉丁方．例如上文的(1)、(2)及此处写出的(3)．这是最大的个数：n 阶的正交拉丁方个数不超过 $n-1$．

1	2	3	4
4	3	1	2
2	1	4	3
3	4	1	2

(3)

现已证明：除了 $n=2$ 和 6 外，对其他 n，都至少有两个正交拉丁方．对指定的 n，正交拉丁方的个数问题至今还远未解决．一个一般的结果是：若 $n=p^k$，其中 p 为素数，则有 $n-1$ 个正交拉丁方．这解决了 $n\leqslant 9$ 的所有情况，也是在实用上最有用的情况．

正交拉丁方的概念，源出于大数学家欧拉(L. Euler)提出的一个问题：有 6 个军种，从每军种中派出上校、中校、少校、上尉、中尉、少尉军官各一名，要把这36名军官排成一个方阵，使每一行及每一列都有各军种的军官一名，各军衔的军官一名．这就是著名的“36军官问题”．不难看出，此问题等价于找两个 6 阶正交拉丁方．直到1900年，才有人证明了此问题无解．

2．正交表

拉丁方只能用于各因素有同一水平的情况，且因素个数至多只能是水平数加 1．正交表的应用则更广．兹举两例来解释这一概念．

试验号＼列号	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

表 $\boldsymbol{A}$: 正交表 $L_8(2^7)$

试验号＼列号	1	2	3	4	5
1	1	2	2	2	2
2	1	1	1	1	1
3	2	2	2	1	1
4	2	1	1	2	2
5	3	1	2	1	2
6	3	2	1	2	1
7	4	1	2	2	1
8	4	2	1	1	2

表 $\boldsymbol{B}$: 正交表 $L_8(4\times2^4)$

上面的表 B 称为正交表 $L_8(4\times2^4)$．L 是正交表记号；8是行数，它表示用此表安排试验时，必须做 8 个处理；“4×2^4”表示表中有 1 列含数字 1、2、3、4，有 4 列含数字 1、2．这表示用此表安排试验时，至多只能容纳 1 个

4 水平因子和 4 个 2 水平因子. $L_8(2^7)$的意义类推.

这种表称为"正交表",因为它有以下两条性质:

(i) 若一列含数字 1、2、…、r, 则每个数字含同样次数(不同的列, r 可以不同).

(ii) 在任一列含同一数字的各位置处, 其他任一列中各数字都有,且含同一次数.

例如,表 B 第 3 列中,数字 1 在第 2、4、6、8 行,而在第 1 列中, 相应位置处数字 1,2,3,4 各占一次. 正是这两个性质,保证了当用这种表作部分实施设计时,能保持平衡性. 我们只举一简单例子说明这表的用法: 设有 4 个因素 A、B、C、D, A为 4 水平, 其余为 2 水平, 全面实施有 $4\cdot 2^3=32$ 次试验. 现作其 1/4 实施,即 8 次. 为此,只须把因素 A 放在表 B 的 1 列处,因素 B、C、D 则可随便占据表上其余 4 列中的 3 列. 例如,B、C、D 分别占第 2、3、4 列. 这一步骤叫"表头设计". 然后,按行读出表头上排有因素的位置的数字, 且按 A、B、C、D 的次序写下来,得(1222),(1111),(2221),(2112), (3121), (3212), (4122), (4211). 这就是排出做试验的那 8 个处理. 表的正交性保证了:任一因素的任一水平的那些处理中,均衡地包含着其余各因素的各水平, 因而在比较时不受它们的影响.

可以证明:正交拉丁方不过是正交表的特例.至于在什么情况下正交表存在, 如何构造出来等问题, 限于篇幅,不能在此多谈了.

三、数据的整理

通过观察或试验得来的原始数据，一般是杂乱无章的，难于从其中直接看出有意义的东西．于是，对原始数据一般尚需要加以整理，以便把我们感兴趣的信息提取出来，并用简明醒目的方式加以表达．整理的方式有二：一是对原始数据进行一定的运算，以算出某些代表性数字，足以反映出数据某些方面的特征．这种数字，在统计学上被称为"统计量"．用数学语言说，统计量就是样本(即数据)的函数．如样本均值*)就是一个常用的重要统计量．二是使用图、表．诸如工厂办公室里挂着的记录逐月生产状况的图表，就属于这一类．一般，设有数据 x_1, x_2, …, x_n，它们都落在区间 (a, b) 内；在 (a, b) 中插入若干分点，把它分为若干份：

$$a = a_0 < a_1 < a_2 < \cdots < a_{l-1} < a_l = b.$$

在应用上常取等分，但也不必非如此不可．对每个区间 $[a_{i-1}, a_i)$，算出 x_1, …, x_n 中落入到这区间里的个数 n_i 及频率 n_i/n，并记下区间的中点，就可以列成一张表：

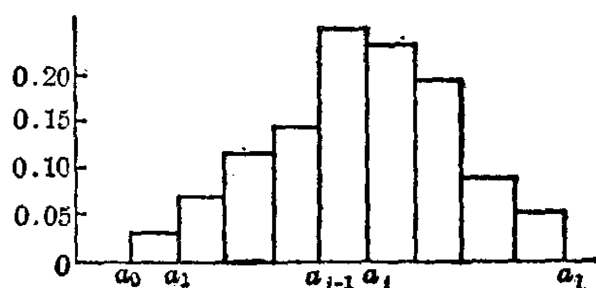
---

*) 在统计学中，习惯上把从样本算出的量冠以"样本"的形容词．如此处的样本均值，及下文的样本中位数，样本方差等．

组 区 间	中 点	频 数	频 率
$a_0 \sim a_1$	$(a_0+a_1)/2$	n_1	n_1/n
$a_1 \sim a_2$	$(a_1+a_2)/2$	n_2	n_2/n
…	…	…	…
$a_{i-1} \sim a_i$	$(a_{i-1}+a_i)/2$	n_i	n_i/n
…	…	…	…
$a_{l-1} \sim a_l$	$(a_{l-1}+a_l)/2$	n_l	n_l/n

这张表可视为是对数据的分布情况的一个粗略的描述. 分组愈多,描述愈精细. 但分组过多,则每组的频数很少,难于看出数据的趋势,失去了整理的原意. 故分组也不宜过多. 一般,当数据较少时,分组数取在10以下;数据较多时,则取在10～20之间.

也可以把这张表转化为一张图:只须在数轴上标出分点 a_0, a_1, …, a_l, 在每个区间上作一个矩形,使其面积等于该区间内的频数或频率即可. 如右图所示. 这种图常称为"直方图".



下文将要介绍的数据的散点图及回归直线,也是通过图形来表示和整理数据的重要方法. 使用统计量和使用图表这两种方法之间有联系:有时,为制作某种图表,需要计算一定的统计量之值(下文的回归直线是一个例子);反之,使用图表有时可以简便地算出所需统计量的(近似)值. 如上表中,数据平均值,即样本均值,近似地为

$\frac{1}{2n}\sum_{i=1}^{l} n_i(a_{i-1}+a_i)$.

严格地说，在统计学理论中，并不把数据整理这一部分作为一个专题或独立的部分去讨论.原因在于，统计理论着眼于统计推断，而对数据作如何的整理，即需要怎样的统计量，要看推断问题的具体形式及所采用的数学模型而定.我们不拘泥于这一点，以便遵循第一节定下的路线来叙述，并借此强调一下统计推断与对数据进行单纯的整理之间的差别.

1. 一维数据的重要统计量

虽说统计量的选择依赖于特定的问题，而在统计实践中使用过的统计量多得不可胜计，但统计学的发展显示，有少数几个统计量有极广泛的应用，经常出现在各种问题中.本节其余部分就主要对这些作一介绍.先考虑一维的情况，即我们只关心总体中每一个体的一个指标值，如人的身高. 若同时也考虑其体重，则数据将是二维的.

一类重要的统计量是用来刻划数据的平均性质的.其中最重要的即我们所熟悉的样本(即数据)均值：设样本为 $x_1, \cdots, x_n$，则 $\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n}$ 就是样本均值. $\bar{x}$ 这个统计量在直觉上为人们所广泛接受，在理论上可以证明它的多方面的优越性，其中之一即曾提到过的，概率论中著名的大数定律.我们提供一个较易理解的依据如下：设

理论上的(即全总体的)平均值为 a, a 未知,通过观察或试验得到样本 x_1, …, x_n, 要由它们算出一个值 b 去估计 a. 由于 x_1, …, x_n 是围绕在真值 a 的附近,我们有理由这样想: b 愈接近 a,偏差平方和 $\sum_{i=1}^{n}(x_i-b)^2$ 应愈小一些 (其所以取平方, 是为防止偏差正负抵消). 易证明 $\sum_{i=1}^{n}(x_i-b)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-b)^2$. 于是只在 $b=\bar{x}$ 时, 偏差平方和才达到最小值. 据此应以 $\bar{x}$ 作为 a 的估计. 这个原则就是著名的"最小二乘法". 此法在科学史上一般都归功于伟大数学家高斯 (C. F. Gauss) 在1799～1809年之间的工作. 这个方法可用于许多问题,在统计学和计算数学上有重要的地位.

除 $\bar{x}$ 外, 刻划平均的另一重要统计量, 是样本中位数. 它定义为 x_1, …, x_n 按大小居于正中的那一个, 或(在 n 为偶数时)正中那两个的平均. 此统计量的直观意义是:数据中超过或低于此值的个数一样多. 在报导中常见到某国某地区处在某条线以下的情况约占一半云云, 即是此统计量的一种应用. 与 $\bar{x}$ 相比, 它的特点在于具有更大的"稳健性", 其含义如下: 当我们收集大量数据时,难免有少数几个发生所谓"过失误差",例如, 小数点打错了地方,而使数据增大或缩小了十倍、百倍等, 这将对 $\bar{x}$ 之值产生较显著的影响, 但对样本中位数则无影响或影响甚微. 有关稳健性的研究,是近年来统计学理论发展的一个方面.

另一类重要的统计量，是为了刻划数据的散布程度．例如，两行业平均工资都是60元，但一个行业内部工资差别很小，而另一个则差别很大．这二者的不同有很大的实际意义，但只看平均值就不能发现．与刻划平均的统计量一样，刻划数据散布程度的统计量很多．其中最重要的是所谓“样本方差”s^2，以及“样本标准差”（也称“样本均方差”）s．s^2 定义为：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

也有用 n 代替上述分母中的 $n-1$ 的．当样本大小 n 较大时，二者差别不重要．s^2 在直观上的意义很清楚：若样本 x_1，…，x_n 的散布比较小，则它们将集中在其平均 $\bar{x}$ 附近，而使 s^2 比较小；反之，则 s^2 将会大．我们看出，这个量的选择，与最小二乘法有关联．与 $\bar{x}$ 一样，在统计学理论中，可证明 s^2 有很多良好的性质．也还有另一些刻划散布程度的统计量，它们在应用上不如 s^2 广泛，但也有其某些特点．

除了平均和散布度这两大类以外，在统计理论和应用上，还有若干常用而重要的统计量．举其中比较易于理解的极值为例．以 $x^*=\max(x_1,\cdots,x_n)$和 $x_*=\min(x_1,\cdots,x_n)$分别记样本 $x_1,\cdots,x_n$ 中的最大者和最小者，它们统称为“极值”．在灾害性现象（地震，水灾等）中，对我们最重要的就是这种极值．如一年中某地各次地震中，震级最大的是多少．另外，如在材料强度试验、可靠性试验中，极值也很重要．日常在报导中，往往听到“这个数字

是五年来的最低点"之类的说法，就是一种极值统计量．由于应用上的重要，"极值统计"已成为统计学中的一个研究题目，有专著问世．另外，$x^* - x_*$ 称为"样本极差"，是刻划数据散布程度的一个统计量．

2．样本协方差，样本相关系数与回归

现考虑多维样本的情况．先设维数等于 2，即对每一个体，我们同时关心它的两个指标．假定有了样本 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)．如果先孤立地取其中的一个指标去看，就分别有一维样本 $(x_1, x_2, \cdots, x_n)$ 和 $(y_1, y_2, \cdots, y_n)$，于是可以计算刻划其平均性质与散布程度的统计量 $\bar{x}$，s_x^2，$\bar{y}$，s_y^2，及其他种种感兴趣的量．这在原则上没有新东西．对我们来说，新东西是与两个指标都有关系的，即刻划两指标之间的关系的那种统计量．这类统计量中，最重要的是样本协方差 s_{xy}，以及与之相联的样本相关系数 r_{xy}．它们的定义是

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$r_{xy} = \frac{s_{xy}}{(s_x\, s_y)}.$$

由著名的许瓦兹(H. A. Schwarz)不等式，得

$$\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2 \leqslant \sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

等号当且仅当存在不同时为 0 的常数 a、b、c，使

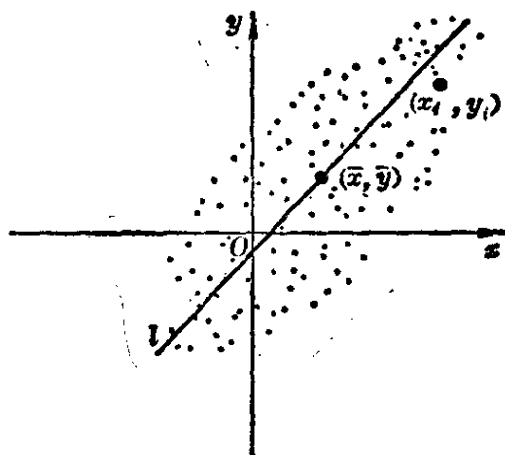
$$ax_i + by_i + c = 0, \quad i = 1, \cdots, n$$

时成立.由此可知,总有 $s_{xy}^2 \leqslant s_x^2 s_y^2$,因而

$$-1 \leqslant r_{xy} \leqslant 1,$$

等号当且仅当 x、y 两指标有严格线性关系时始成立.由于这个原因,有时也把 r_{xy} 称为线性相关系数:$|r_{xy}|$ 愈接近于1,x、y 之间线性关系的程度愈大. r_{xy} 的符号则显示相关的方向:$r_{xy}>0$ 时称为正相关,$r_{xy}<0$ 时称为负相关. 这些,我们现在从另一不同的,十分重要的角度来说明.

在 (x, y) 平面上取一个直角坐标系;把每个样本 (x_i, y_i) 标在这坐标平面上得一个点.若样本大小为 n,则得到由 n 个点构成的一张图(如右图),称为"散点图".作散点图,在某种意义上是一种最重要的整理数据的方法.因为一看这张图,对两指标值的平均、散布与其关系的大致情况,心中就有了一个概念. 这其中最重要的,就是帮助我们探索两指标之间的关系. 如图,我们看出 x、y 之间有一定线性关系的趋势,但又不严格为线性.我们想要找一条直线(图中的 l),能大体上反映这个趋势.这条直线 l 本身,就是对全部数据的一个形象化的概括,但有两个问题:(i) 这条直线如何找? (ii)其代表性如何? 我们下面将看到:这两个问题的回答,都与线性相关系数 r_{xy} 有关.



为找直线 l,使用最小二乘法. 我们想要找一条直线

$y=a+bx$，在某种意义上与散点图上各点尽可能接近．按这个直线方程，当 $x=x_i$ 时，y 应为 $\hat{y}_i=a+bx_i$，但实际观察结果为 y_i，偏差平方和为

$$L(a,b)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2.$$

要找 a、b，使 $L(a,b)$ 最小．用微积分求极值法，或用简单的配方法，不难求得，只在

$$b=\hat{b}=\frac{r_{xy}s_y}{s_x},$$

$$a=\hat{a}=\bar{y}-\hat{b}\bar{x}\left(=\bar{y}-\frac{r_{xy}s_y\bar{x}}{s_x}\right)$$

时，$L(a,b)$ 达到最小值．这条直线 $y=\hat{a}+\hat{b}x$ 称为数据的(经验)回归直线（“经验”的意思，表示它系由数据得来）．我们注意到，它通过散点图的中心 $(\bar{x},\bar{y})$．其代表性如何，则要看偏差平方和 $L(\hat{a},\hat{b})$ 的大小．此值愈小，代表性愈大．通过简单的初等代数计算，不难得到

$$L(\hat{a},\hat{b})=\sum_{i=1}^{n}(y_i-\bar{y})^2(1-r_{xy}^2)$$

$$=(n-1)s_y^2(1-r_{xy}^2).$$

利用这些公式，就可以对相关系数 r_{xy} 的意义，给予更清楚的解释．首先，$\hat{b}$ 的符号与 r_{xy} 同．若 $r_{xy}>0$，则 $\hat{b}>0$，而直线 l 的趋势是：当 x 增加(下降)时，y 随之增加(下降)．这是我们当 $r_{xy}>0$ 时把 x、y 的关系称为“正相关”的理由．类似地，得到“负相关”的解释．

其次，$\sum_{i=1}^{n}(y_i-\bar{y})^2$ 即 $(n-1)s_y^2$，反映了 y 数据的散

布程度；而 $L(\hat{a}, \hat{b})$则反映了在指标 y 中消除掉 x 的影响后，所剩余下的或残留的散布度(以此之故，$L(\hat{a},\hat{b})$常称为残差平方和)．这表明：在 y 的全部散布度中，因 x 的影响所占比率为 $100r_{xy}^2\%$．$|r_{xy}|$愈接近 1，这比率愈大；当 $r_{xy}=0$ 时，

$$L(\hat{a}, \hat{b})=\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

表示 x 对 y 无影响，这时，称 x、y“不相关”．通过这一分析，看出 r_{xy} 是衡量 x、y 之间的线性关系的良好指标．

经验回归方程 $y=\hat{a}+\hat{b}x$ 的斜率 $\hat{b}$，常称为(经验)线性回归系数．在实用上，回归方程常用于预测(即已知x值时，预测相应的 y 值)．所用预测量就是 $\hat{a}+\hat{b}x$．这预测量的好坏，当然取决于回归方程的代表性，即 $L(\hat{a}, \hat{b})$ 的大小．这种预测问题在应用上很常见，因而回归是统计方法中的一个极重要的方法．形式上，我们也可以掉转 x、y 的位置．但在具体问题中，x、y 中谁为预测者，谁为被预测者，要看实际情况．例如，只能由施肥量预测产量，而不能反过来．

在历史上，回归这个名词是英国著名统计学家兼生物学家高尔顿(F. Galton) 在上世纪八十年代提出来的．他考察了1078对夫妇，以夫妇身高的平均作为 x，其一成年儿子的身高作为 y，描出1078组数据的散点图．当 x 增加时，y有增加的趋势．然而，高尔顿注意到下述有趣的现象：父代平均身高为 68 英寸，子代则为 69 英寸．依此，一

般人会预期：当父代身高固定在某值 x 时，子代平均身高应为 $x+1$ 左右.但实际不然.例如,当 $x=72$ (大于平均值 68)时,y 的平均值只有 71,虽则大于总平均 69,但比 x 还小一些.反之,若 $x=64$(小于平均值 68),则 y 之平均值为 67,虽则低于子代总平均 69,但比$x+1$(即 65)还大.这意味着,在本例中子代身高平均值有回归于其中心(69)的倾向.以此之故，高尔顿把本例中 x、y 的关系加上"回归"的称呼. 然而,这只是在本例中特有的现象,并不是有普遍性的特征. 把它作为变量之间的关系的称呼并不恰当.只是这名称现在已经成了习惯,无法加以改变了.

一般地，可以考虑多个变量之间的关系，也可以不限于线性关系，这些都属于统计学中"回归分析"这分支. 这是一个在应用上极重要，在理论上也很发展的分支.

如果考虑的指标个数多于2,则情况也相似. 每个指标的观察值构成一个一维样本,可计算其样本均值、样本方差等,这没有新的东西.有兴趣的是那些反映指标之间关系的统计量,主要的仍是样本协方差与相关系数.如果有 k 个指标(样本是 k 维的),则有

$$C_k^2=\frac{k(k-1)}{2}$$

个样本协方差和相关系数. 自然，也可以考虑多于两个指标的关系问题. 这一般只涉及样本均值、方差与协方差.

四、统 计 推 断

统计推断是数理统计学理论的主要部分．现行的统计推断理论，是建筑在概率论的基础上的，因此，本节要求读者了解概率论的一些初步知识．

前已说过，统计推断，就是根据从总体中抽出的样本，去推断总体的性质．由于我们关心的总是总体中的个体的某项指标，所谓总体的性质，无非就是这些指标值的集体的性质，而概率分布正是刻划这种集体性质的适当工具．因此，在理论上可以把总体与概率分布等同起来．例如，当指标值的概率分布为正态分布时，我们可称这个总体为正态总体，等等．

如果指标值的概率分布完全已知，则从统计学的观点看，样本已无用武之地——没有什么需要藉助于样本去推断的东西．总体的性质包含在其概率分布中，只有当这种分布中包含未知的成份时，才发生推断问题．

例如，有理由假定：在一大群人中，身高服从正态分布 $N(a,\sigma^2)$．均值 a 反映这群人的平均身高，而方差 σ^2 则反映身高的不均匀程度．我们虽可假定身高服从正态分布，但 a 和 σ^2 这两个参数则不知道，它们是指标（身高）的概率分布中的未知成份，即推断的对象．又如，大批生产的一种电子元件，在一定条件下，有理由假定元件

寿命(我们关心的指标)的概率分布为指数分布,其概率密度为

$$f_\theta(x)=0,\text{ 当 } x<0;\quad f_\theta(x)=\frac{1}{\theta}e^{-x/\theta},\text{当 } x\geqslant 0.$$

参数 $\theta>0$ 是这个分布的未知成份,它就是元件的平均寿命,这正是应用上有兴趣的量,而成为统计推断的对象.

这样,我们就可以一般地把统计推断的问题,抽象为如下的数学模型:总体的概率分布 $F_\theta(x)$包含了其值未知的参数 θ(这里,θ 可以是向量,如在正态总体中有 $\theta=(a,\ \sigma^2)$).从该总体随机抽样,得样本 $x_1,\ \cdots,\ x_n$,要通过后者,去获得对 θ 的某些了解.这后一点的确切含义,依赖于所要回答的问题的性质.主要的形式有两种:

1. 估计问题. 即要通过样本 $x_1,\ \cdots,\ x_n$ 对 θ 的值作出估计.如估计上述指数分布的参数 θ. 这问题的实际含义无非是:从一大批电子元件中抽出 n 件,测得其寿命为 $x_1,\ \cdots,\ x_n$,要利用这些数据,去估计整批元件的平均寿命.由于估计的对象是参数,常称为参数估计. 参数估计又分为两种基本形式:点估计和区间估计.前者是用一个数值作为未知参数 θ 的估计值,后者则用一个区间,把 θ 估计在这个区间内.犹之如估计某人的年龄为25岁,是点估计;估计其年龄在20～30岁之间,是区间估计. 在统计上,点估计就是样本 $x_1,\ \cdots,\ x_n$ 的一个函数 $\hat{\theta}(x_1,\ \cdots,\ x_n)$(即统计量),称为 θ 的一个“估计量”. 每有了样本 $x_1,\cdots,\ x_n$,即可代入其中而算出具体数值 $\hat{\theta}$,用以估计 θ.人们常称由估计量算出的具体数值为“估计值”.至

于区间估计，则不过是两个统计量 $\hat{\theta}_i(x_1, \cdots, x_n)$，$i=1$、2，满足条件 $\hat{\theta}_1 \leqslant \hat{\theta}_2$. 每有了样本 $x_1, \cdots, x_n$，就代入其中算出具体数值 $\hat{\theta}_1$、$\hat{\theta}_2$，而将 θ 估计在区间$[\hat{\theta}_1, \hat{\theta}_2]$之内.

例如，用样本均值 $\bar{x}$ 估计总体均值（用我们现在的说法，就是总体的概率分布的均值），是一个常用方法. 故$\bar{x}$可用于估计正态分布 $N(a,\sigma^2)$中的a，指数分布中的 θ，等等. 停留在数据整理这个角度上，人们会觉得算术平均 $\bar{x}$ 是个"天然合理"的量，没什么值得进一步讨论的东西. 但从统计推断的角度去看，则可以提出很多问题. 主要是这估计的精度如何？比如问：$\bar{x}$ 与总体均值的误差不超过1 的可能性(概率）有多大？这要求在正态分布的假定下算出$P(|\bar{x}-a|\leqslant 1)$，在指数分布的假定下算出$P(|\bar{x}-\theta|\leqslant 1)$，等等. 由于总体分布的假定不同，这概率的算法及其值都不一样. 又，这估计量的精度还可以通过其均方误差($E(\bar{x}-a)^2$, $E(\bar{x}-\theta)^2$ 等)表现出来. 对上述两个分布，均方误差可分别算出为 σ^2/n(对 $N(a,\sigma^2)$)以及 θ^2/n(对指数分布). 是否可以找到均方误差比这更小的估计量？这就是一个不易回答的理论问题. 对上述两个分布而言，可证明这样的估计量不存在. 但是，若总体的概率分布是区间$(0, \theta)$上的均匀分布 $R(0,\theta)$，即有概率密度

$$f_\theta(x)=\frac{1}{\theta}, \text{ 当 } 0<x<\theta;$$

$$f_\theta(x)=0, \quad \text{对其他 } x, \ (\theta>0)$$

则可以证明这样的估计量存在. 如 $\frac{n+1}{2n}\max(x_1, \cdots, x_n)$ 即为其一. 又，均方误差不过是可能提出的优良性

准则中的一种，因此，就可以提出问题：在另外的优良性准则下，$\bar{x}$ 这个估计量的表现如何？从这个简单例子，可以看出点估计理论的丰富内容的简单轮廓．正如我们在前面曾指出过的：宣布把样本均值 $\bar{x}$ 作为总体平均的估计，表面上好像只是形式地跨出了一步，却是一件不简单的事情，需要许多理论上的论证来支持．

区间估计的优良性可分两个方面去考察．一方面是可靠度，即区间 $[\hat{\theta}_1(x_1,\cdots,x_n),\ \hat{\theta}_2(x_1,\cdots,x_n)]$ 能包含未知参数 θ 的可能性多大，就是概率 $P\{\hat{\theta}_1(x_1,\cdots,x_n)\leqslant\theta\leqslant\hat{\theta}_2(x_1,\cdots,x_n)\}$，它称为区间估计 $[\hat{\theta}_1,\ \hat{\theta}_2]$ 的"置信系数"；另一方面是精度，区间愈短，精度愈高——当然，精度也可以不直接通过区间长度去衡量．我们希望找到这样的区间估计，其置信系数尽量接近 1，而区间长度尽可能小，可是这两者有矛盾．犹之如你要把一个人的年龄估计在一个很小的范围内，你就要冒比较大的出错的风险．在统计学上，处理这问题的作法是按照在本世纪三十年代开创了区间估计理论的奈曼(J. Neyman) 的方案，即在保证一定的置信系数的前提下，使精度尽可能高．在统计学上，常用0.95，0.99，0.90这些数，尤其是0.95．区间估计的一个最重要的例子，是为估计正态分布 $N(a,\sigma^2)$ 的均值 a 的所谓"t 区间估计"：设 $x_1,\cdots,x_n$ 是从正态总体 $N(a,\sigma^2)$ 中抽出的样本，以 $\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n}$ 和 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 记样本均值和方差，并给定置信系数

为0.95，则 a 的区间估计可取为

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1}(0.05),$$

即

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{n-1}(0.05),\ \bar{x} + \frac{s}{\sqrt{n}} t_{n-1}(0.05)\right].$$

这里，n 是样本大小；至于 $t_{n-1}(0.05)$，那是一个根据所谓"t 分布"、样本大小以及所给的置信系数0.95 $(0.05 = 1-0.95)$而确定的值，有表可查，例如

k	2	3	4	5	6	7	8	9	10	∞
$t_k(0.05)$	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228	1.960

考察一组具体数据：设从一大群人(其身高可认为服从正态分布 $N(a, \sigma^2)$)中抽出 8 个，量得其身高为(厘米)：

168，172，170，180，176，174，165，173，

算出 $\bar{x} = 172.25$，$s = 4.683$. 又 $n-1=7$，而 $t_7(0.05) = 2.365$. 于是作为点估计，我们用 $\bar{x} = 172.25$ 去估计这群人的平均身高 a，作为区间估计，则用

$$172.25 \pm \frac{4.683}{\sqrt{8}} 2.365 = [168.33,\ 176.17],$$

其可靠性为0.95. 如果要把可靠性提高到0.99，则应在上述公式中，把 $t_{n-1}(0.05)$ 改为 $t_{n-1}(0.01)$. 对 $n=8$，此值为 3.499，而区间估计将变为[166.46，178.04]，精度比原来的差了.

即使在这比较简单的情况，仍能提出很多问题. 总的说，就是上述 t 区间估计在种种准则下的优良性问题.

其中,有的直到目前仍是研究的对象.

2. 检验问题. 先看一个例子.设某工厂生产了一大批电子元件. 其寿命可假定服从指数分布, 参数 θ 的值(即整批产品的平均寿命)是未知的. 现设这批产品的使用者立下了一个界限:只有在平均寿命 θ 不小于某个值 θ_0 (如 $\theta_0=5000$ 小时)时,才接受这批产品. 为此,从这批电子元件中抽出若干个,测得其寿命为 $x_1, \cdots, x_n$; 要据此判断"$\theta \geqslant \theta_0$"是否成立. 又如,一种产品中所含杂质的量,可假定为服从正态分布 $N(a, \sigma^2)$. a 是杂质的平均含量,现使用者要求这平均含量 a 不超过某个界限a_0. 于是,抽出 n 个样品,测得其杂质含量为 $x_1, \cdots, x_n$, 要由此判断"$a \leqslant a_0$"是否成立.

由此看到,在这一类问题中, 我们有一个待判断的,与总体概率分布的参数有关的命题: $\theta \geqslant \theta_0$, $a \leqslant a_0$, 等等. 在统计学上,把这称为"假设".使用样本去判断一个假设是否成立,称为"假设检验",它是与参数估计并列的、统计推断的两种基本形式之一.

假设检验问题的具体回答只有两种:接受假设,或否定假设. 问题是要建立一个法则,使当一有了样本时,这个法则就能决定是接受还是否定假设. 任何一个这样的法则,都叫做所给假设的一个"检验".

举例来说,在前面所提"$a \leqslant a_0$"这个假设的检验问题中,我们先用 $\bar{x}$ 估计 a. 由于 $\bar{x}$ 与 a 接近,我们在直观上就易于接受下述作法:应当在 $\bar{x}$ 较小时,比方说 $\bar{x} \leqslant c$时,接受假设 $a \leqslant a_0$; 若 $\bar{x} > c$, 则否定 $a \leqslant a_0$. c 不一定即取

为 a_0,因为 $\bar{x}$ 只是 a 的估计值,而不就是 a.那么,c 取多少为好呢? 这就涉及理论上的问题. 根据奈曼和皮尔逊(E. S. Pearson)在本世纪二、三十年代所发展的理论,是用如下的方式来处理这个问题:他们指出,在检验一个假设时可能犯两种错误之一,一是假设本来对,但被否定了(第一种错误); 一是假设本来不对,但被接受了(第二种错误); 他们提出,控制第一种错误的概率, 使之不超过给定的数 α(这里,α 一般很小,如 0.05,0.01,0.10等,称为检验的水平); 在选择了一定的检验统计量(如上例的 $\bar{x}$)后,根据给定的水平 α 及检验统计量的分布, 就可以决定界限 c. 但是,对一个假设来说,可用的检验统计量很多.如在上例,也可用样本中位数 m,而当 $m \leqslant c'$ 时接受假设 $a \leqslant a_0$. 在众多可能的检验统计量中怎样选择其一呢? 奈曼和皮尔逊提出的原则是: 在控制第一种错误概率的前提下,使第二种错误的概率尽可能小. 而这"尽可能小"一语,又有形形色色的解释, 相应于种种具体的准则. 在这个原则之下,发展了一整套假设检验理论,成为现在统计推断理论的一个重要构成部分.

除了这两种基本的推断形式以外, 另有一种常用的推断形式,可说是介于两者之间:参数估计可能的推断结果是无穷多个,假设检验只有 2 个. 而在有些问题中,可能的推断结果多于2,但个数为有限.例如,有 5 个玉米品种,要挑选其中产量最高者.如果我们假定这 5 个品种的产量分别服从正态分布 $N(a_i, \sigma^2)$, $i=1, \cdots, 5$,则问题在于找出一个 i, 使 a_i 在 $a_1, \cdots, a_5$ 中最大. 这时, 可

能的推断结果有 5 个.

参数估计和假设检验不仅是两种基本的推断形式，它们各自也构成统计学中的基本分支. 在统计学中存在着一些学科分支，其中的两个——抽样技术与试验设计，是与获取数据有关；其他分支则主要关系到统计推断的，重要的如多元统计分析，它是讨论多维数据的统计推断的，相关回归分析，则是讨论带随机性的变量的关系的分支，它也可视为多元分析的一部分，等等. 另有一些分支，则是联系某种特殊应用而建立的. 还存在统计学与其他学科的一些边缘性分支学科，如生物统计学、数量遗传学、计量经济学等.

五、统计学的应用

在前面的叙述中，我们已提到过统计学的某些应用. 为使读者对这门学科的重要实际意义有更深刻的印象，我们再花点篇幅来作一点较系统的介绍.

统计学最主要的应用领域有二：工农业生产和社会、经济领域. 田间试验的适当的设计及统计分析，是统计方法在农业中应用的直接形式. 其实，我们在第三节中介绍的试验设计的基本思想和方法，最初就是从田间试验开始发展起来的. 农业生产中有许多可变因素，像种子品种，播种量，肥料和农药的种类及数量，耕作方法及田

间管理方式的选择，等等．为提高产品的数量及质量，需要通过试验，对这些因素在允许的范围内进行选择．农业试验有其特点，即试验周期长(因而组织一次试验不易)，环境因素变异大．在这种不利的条件下，如不对试验的设计安排作精心考虑，并使用有力的统计分析方法，则得不出什么有用的结论．这一点解释了为什么试验设计及其统计分析的发展，始自农业和生物方面的应用．统计方法在农业上还有一些较为间接的应用．例如，培育优良品种是农业上的一个重要问题．在学科上说，这种问题属于数量遗传学研究的范围．而其中就使用了大量的统计方法．如在遗传力的计算上，使用了很复杂的回归分析和方差分析的方法．

统计方法在工业上的应用，比其在农业、生物上的应用开始得略晚．最初，在本世纪二十年代后期，有人开始把统计方法用于成批产品的抽样验收，以及生产过程中的工序控制．稍后，在三、四十年代，又有人把在农业中发展的一套试验设计的思想和方法用于工业领域，并有所发展．例如，在农业试验设计中，部分实施(见第三节)不常用，而在工业试验中则极为常见．不过，统计方法在工业中的大量应用并取得引人注目的成效，是二次大战以后的事情．这一点当然不仅与统计学本身的发展有关，更重要的是与战后时期工业的飞速发展有关．二者起了相互促进的作用．大略言之，统计方法在工业上的应用主要有两个方面．一个方面是试验的设计及其统计分析．在试制新产品、改革工艺流程、使用代用原材料及寻求适当的

配方等问题中，都需要通过试验，去决定在大量的影响产品质量和数量的因素中，那些是主要的，那些是次要的，并决定一组优良的生产条件．正交设计、回归设计与分析、方差分析、多元分析等，都是处理这个问题的有效工具．另一类应用可总结在"统计质量管理"这个名目下．它是用统计方法，对工业生产过程中及事后的验收和使用中，对产品质量进行评估和控制．如产品抽样验收，是根据从一大批产品中抽出一小部分作检验，以判定该批产品可否接受．从理论上说，这不过是一个假设检验问题．但由于其在应用上的重要性，目前关于这方面已出现了好几本专著．不少国家的有关部门，包括武装部队在内，都编制了特定的标准，做这项工作离不开统计学理论的指导．统计质量管理的另一个重要内容是工序控制，即在产品制造过程中，通过抽查，发现生产过程可能超出控制范围的一些统计方法．这些方法在学理上也不过是关于几个常见分布（正态分布、二项分布等）的检验问题．但一经与应用结合，就有了丰富的内容．另外，可靠性统计分析也是统计质量管理的一个重要方面．例如，一部复杂的装置由大量的元件组成，当这些元件中的一个或某些个不能正常工作（失效）时，该装置就不能正常工作；而元件在何时失效是随机的，因此，整部装置的可靠性可以用概率论的方法去计算，并用统计学的方法进行估计．

以上这些统计方法，在战后时期，在一些工业发达的国家中逐渐得到了普遍的应用，获得了良好的经济效益．一个有代表性的例子是日本．有人估计，在日本战后高速

经济增长中，有5%的份额可归功于统计方法的使用．这个比率的准确性如何姑且不谈，但无可怀疑的是，统计方法的使用对日本经济的发展确实起了重大作用．一个例证是：日本在这方面的成就，包括它对从西方“输入”的一些统计方法的改造（改造的目的是使之有更便于使用的形式，以便有更多的人，包括具有一定文化水平的工人，都能使用它），得到了像美国这样工业和统计都很发达的国家的重视．

我国在这方面的起步较晚．五十年代末期，开始在小范围内做了若干工作．近几年来，这方面努力的步伐加快了，但目前与先进国家比仍有不小的差距．

从现代统计学发展的早期直到现在，统计方法在社会、经济领域中的应用，都在其全部应用中占很大的比例．有资料表明，在统计学发达的国家中，统计学家就业人数的比例，以这个领域为最高．从性质上说，在这个领域内的应用可以分成两类．一类是单纯的抽样调查性质的．即为了要了解一个包含极大数目的个体的总体的情况，而从其中按一定的方式抽出一些个体作调查．这么做的原因、做法及其优点等，已在前面第二节中作了充分介绍．这是一个很大的应用领域．在不少这类性质的应用中，人们事先并未对总体的状况形成什么看法，也不需要通过抽样去验证(有关于总体的)某种理论，而纯粹是为了“了解情况”．我们把这类应用称为单纯的抽样调查．一般说，这类应用的困难在于目标的适当确定及抽样的组织工作，而在统计理论方面则较为简单．另一类的应用往

往也涉及抽样调查，但要求更深刻一些．比如说，通过抽样调查的数据去探索某种理论上的规律性，或验证所提出的某种规律性是否与实际符合等．举几个例子．制定了某种人口政策，要探索在这种政策下，人口将以怎样的规律变化，或在制定政策时，预计人口将以某种规律增长．在经过一段时间后，通过抽样调查，去验证实际情况与最初的设想是否符合，应作何修正．又如，在资本主义国家中，政府所实施的金融以及经济、社会政策，对具有盲目性的市场经济有很大的影响；这种影响确切的情况如何，是否沿着政策制定者所希望的方向去发展，都要通过收集数据进行分析，相当大的程度上是统计分析．还有，下面这个例子也可以概括一个方面的应用：1927年，美国心理学家斯彼尔曼曾在其一本著作中提出一个假说：一个人在某方面的智力，由两个因素组成，一是其"一般智力"，一是与此特定方面有关的智力因素．这不是一个在学理上可以严格证明或否定的命题．人们对这个说法的态度，也容易囿于自身的狭隘经验．只有通过适当的试验，并进行统计分析，才是解决这个问题的正确途径．加州大学的特利昂教授用老鼠作了这样的试验，他得出的结论是否定的．虽然这还不能与对人类的试验等同起来，但有相当的参考价值．不难想像，这类性质的问题在社会领域中是很多的，其解决必然用到统计分析方法．

总的说，近年来特别在西方，社会研究定量化的趋势愈来愈明显．至于在经济科学中，由于其性质，定量化的趋势比其他社会科学部门更早，且程度更深．如早在二、

三十年代，时间序列分析方法就已用于市场预测．现在已建立了一门边缘性质的学科——数量经济学，其中使用了许多近代数学的知识，包括概率论和统计学，从简单的回归分析方法到艰深的随机过程统计方法，都在其中找到了应用．

除了上述这两个主要方面的应用外，如前面曾指出的，统计方法在几乎人类活动的一切领域内，都或多或少能发挥一些作用．例如，医学是较早使用统计方法的一个重要领域．我们经常在各种书籍及报纸杂志上读到，某某因素是导致某种疾病的一个原因，如吸烟使患癌症的危险性增加，饮酒过量对肝脏有损害，而适量饮酒则可能有益于健康，吃盐过多对健康有多方面的危害等(如导致高血压)，这些大多是首先通过统计分析而发现的，然后促使学者们对其机理进行研究．有的也可能是从纯学理的分析提出来，但也必须寻求统计资料的验证．另外，一种药物对治疗某种疾病是否有效，效果多大，几种药物或治疗方法效果的比较，最后都必须诉诸临床试验，用统计分析的方法确定．这是因为人群的变异性很大，同患一种病，因体质、年龄、遗传基础以至以往的生活史和健康史等等方面的差异，对同一种药物或治疗方法的反应就会有差异．只有在精心设计、进行大量观察，并使用正确的统计方法去进行分析，才能得出科学上站得住脚的结论．有时，人们从广告中看到某种药物治疗某病的有效率很高(百分之九十或更高)，而实际使用效果却并不理想，这只要看看试验规模多大，样本如何收集，以及数据的统计

分析是如何做的,就不难从其发现问题.

统计方法在自然科学和技术科学中的应用，少量的是属于纯学理的，而大量的则是直接应用的性质——解决人们在面向自然的种种实践活动中所碰到的问题．当然，这两个方面并非截然分开的，可能在某项研究工作中兼有这两方面的目的.属于前一方面的一个典型例子，是本文前面提到的孟德尔(J. G. Mendal)遗传定律．其实,一般地讲,自然科学(数学除外,通常并不把数学看作为自然科学的一个部门)中,任何规律性都有一个经受统计检验的问题．例如，用适量的观测数据对开普勒 (J. Kepler) 行星运动定律进行统计检验，可以认为是符合的；但如用极大量的观测数据去检验，则会发现其符合程度并不佳.因此,就弄清楚了：开普勒的行星运动定律只是在一定的误差限度内正确，而这自然与牛顿力学的近似性质有关.至于在应用性的研究中,常因对所研究的现象的规律性认识不充分，而不能不在很大程度上通过对试验和观察数据的分析,建立一些经验性的规律(如经验公式),并利用它去处理所面临的问题.如在地震预报、地质探矿和气象预报中,统计方法都有很多应用.像在地震预报的研究中,人们通过用统计方法分析以往的资料,可能会发现某级以上的大地震的发生，存在着种种可能的周期.人们无法从学理上严格证明这种周期的存在,它可能只是一个很粗略的近似，但毕竟是认识上的一种进步,且有实际意义.又如我国统计学者在使用统计方法找矿这一方面，作出了一些很有实际意义的成果．这种工

作也可能提出对所研究的现象的规律性的认识.

六、简单的历史与现状

在本节中,我们将简略地回顾一下统计学的发展史.包括发展过程中所经历的一些大事,以及对这门学科的创立与推进有特别重大影响的那些学者的贡献.由于我们不能涉及本学科过多的细节内容,所作的介绍只能是很粗线条的.我们也准备对我国统计学的状况作一些介绍.另外,我们也想顺便谈谈统计学的目前状况和有待解决的一些重大问题,供对这方面有兴趣的读者参考.

在我国历史典籍《二十四史》中,有不少钱粮户口、水灾地震等有关国情的记载.这是统计性质的工作,当然还不能算作是现代意义下的统计学,因为这只是有关事实的记录、整理,而没有在一定的理论的指导下,作出超越数据范围之外的推断.现代统计学的产生,一方面是由于在各种领域内应用上的需要;一方面由于近代数学和概率论的发展,提供了把一些多少是从经验上提出的,个别的方法,加以理论上的提高和系统化.

高斯(C. F. Gauss)从描述天文观测的误差而引进正态分布,并使用最小二乘法作为一种估计方法,是近代数理统计学发展初期的重大事件. 18 世纪末到 19 世纪初期的这些贡献,有很大的影响. 例如,用正态分布描述观

测数据后来被广泛地用到生物学中；其应用是如此普遍，以致在上世纪相当长的时期内，包括高尔顿在内的一些学者，认为这个分布可用于描述几乎是一切常见的数据．直到现在，有关正态分布的统计方法，仍占据着常用统计方法中很重要的一部分．最小二乘法方面的工作，在本世纪初以来又经过了一些学者的发展，如今成了数理统计学中的重要方法．

从高斯到本世纪初这一段时间，统计学理论发展不快．但仍有若干工作对后世产生了很大的影响．其中，如贝叶斯(T. Bayes)在1763年发表的《论有关机遇问题的求解》，提出了进行统计推断的方法论方面的一种见解，在这个时期中逐步发展成统计学中的贝叶斯学派(如今，这个学派的影响愈来愈大)．再如前面提到的高尔顿在回归方面的先驱性的工作，也是这个时期中的重要发展．

数理统计学发展的第二个阶段，是从上世纪末期到二次大战结束．现在，多数人倾向于把现代数理统计学的起点和达到成熟定为这时期的始末，因此，这是数理统计学发展史上极重要的一个时期．

这确是数理统计学蓬勃发展的一个时期，许多重要的基本观点、方法，统计学中主要的分支学科，都是在这个时期建立和发展起来的．以费歇耳 (R. A. Fisher, 1890～1962)和卡·皮尔逊 (K. Pearson, 1856～1936)为首的英国统计学派，在这个时期起了主导的作用，特别是费歇耳．

卡·皮尔逊发现，有不少生物学方面的数据有显著

的偏态，不适合用正态分布去刻划．为此，他提出了一个后来以他的名字命名的分布族．为估计这分布族中的参数，他提出了“矩法”．为考察实际数据与这族分布的拟合优劣问题，他引进了著名的“χ^2检验”，并在理论上研究了其性质．这两方面的工作，对统计学的应用及以后的理论发展，都有重要的意义．

费歇耳对数理统计学的发展作出了最大的贡献．在此，我们只能列举他的几项主要工作：

1．参数估计方面．他提出了著名的“极大似然估计法”．这是应用上最广的一种估计方法．他在二十年代的工作，奠定了参数估计的理论基础．

2．试验设计与方差分析．我们在第2节中叙述的试验设计方面的内容，包括设计的三大原则，是费歇耳及其合作者叶茨(F. Yates)所开创的．他们还发展了分析这种试验数据的统计方法——方差分析法．

3．多元分析、相关回归．费歇耳系统地研究了正态分布样本的一些重要统计量的抽样分布，这些都是多元分析、相关回归等分支的奠基性工作．

4．其他．费歇耳在假设检验和一般的统计思想方面，也都作出过重要的贡献，后者包括他提出的一种新的统计推断思想——信任推断法．

在这个时期作出了重要贡献的统计学家中，还应当提到奈曼和依·皮尔逊．他俩人联合发展了假设检验的系统理论．奈曼还发展了区间估计的理论．他们工作的要旨，曾在第4节中介绍过．1946年，瑞典统计学家克拉美

(H. Cramer)的《统计学数学方法》一书问世．这是第一部严谨而较系统的数理统计学著作，其中总结了上文提到的主要成就.可以认为，这本著作的问世，标志了数理统计学已成为一门成熟的学科.

从战后到现在，是统计学发展的第三个时期.这是一个在前一段发展的基础上，随着生产和科技的普遍进步，而使这个学科得到飞速发展的一个时期.同时，也出现了不少有待解决的大问题.为节省篇幅，我们把这一个时期的发展，总结为以下四个方面：

一是在应用上愈来愈广泛．统计学的发展，一开始就是应实际的要求，并与实际密切结合的．在二次大战前，已在生物、农业、医学、社会、经济等方面有不少应用，在工业和科技这方面也有一些应用．而后一方面在战后得到了特别引人注目的进展．例如，归纳到"统计质量管理"名目下的众多的统计方法，在大规模工业生产中的应用取得了很大的成功，目前已被认为是不可缺少的．我们在前面已谈到过这些方面对日本在战后的经济发展中的作用．在其他国家中也取得了成效．统计学应用的广泛，也可以从下述情况得到印证：统计学已成为高等学校中许多专业必修的内容，统计学专业的毕业生的人数，以及从事统计学的应用、教学和研究工作的人数的大幅度增长，有关统计学的著作和期刊杂志的数量的显著增长．如在美国，每年统计学方面毕业的大学生人数，与数学方面的大学毕业生人数相当或略多．从事统计学方面的工作者已有十余万人，每年出版统计学方面的著作和教科

书数百种，统计学方面的专业杂志就有四、五种，部分有关的还不计在内．

二是统计学理论——数理统计学方面也取得了重大的进展．虽然可以说，在1946年克拉美完成他的著作《统计学数学方法》时，数理统计学可算是有了完整的体系．但许多方面的研究还只是很初步的，甚至没有开始．现在则面貌大为改观了．理论上的成就，综合起来大致有两个主要方面．一个方面与瓦尔德(A. Wald)所提出的“统计决策理论”有关．这方面留待下面再谈．另一个方面就是大样本理论，即在样本大小无限增加时，统计量与统计方法的极限性质的理论．不过，随着这种理论的纵深发展，也就产生了一个重要问题：有的学者认为，纯理论方面的发展，使统计学发展初期(指战前时期)与实际密切结合的传统有所削弱．甚至认为这是一个“危机”．就大样本理论来说，确实有些成果在数学上很深刻和精细，但已没有多大实用价值．这是因为，在实际问题中，样本大小总是有限的．对某一具体的样本大小而言，极限结果的误差多大，缺乏有用的估计．故有人认为，发展有实用价值的大样本理论，是目前数理统计学所面临的一个重要课题．

三是电子计算机的应用对统计学的影响．这主要在于以下几个方面：首先，一些需要大量计算的统计方法，过去因计算工具不行而无法使用；有了电子计算机，这一切都不成问题．前面提到过，在战后，统计学应用愈来愈广泛，这在相当程度上要归功于计算机．特别是对于高维数据的情况．对这种情况，传统的统计理论中提供的模

型(如多维正态模型),一般不甚符合实际.有的学者发展了一些思想和方法,着重在利用计算机在短时间内处理大量数据的能力,以直接从数据出发探索可用的模型,以及有效地提取数据中的信息.对这一方面,有的学者寄予很大的希望,认为是未来统计学发展的方向之一.目前,在这方面已出现了某些受到注意的工作,但在这方面能走多远,还要拭目以待.

电子计算机的使用,对统计学的另一方面的影响是:按传统的数理统计学理论,一个统计方法的效果如何,甚至一个统计方法如何付诸实施,都有赖于决定某些统计量的分布,而这常是极困难的.数理统计学家往往只好退而求其次——转向大样本方法.而这样做,又有前面所指出的困难.现有了计算机,就提供了一个新的途径:模拟.例如,用模拟的方法去决定某个抽样分布的分位点,很容易达到实用上满意的解决.为了把一个统计方法与其他方法比较,可以选择若干组在应用上有代表性的条件,在这些条件下,通过模拟去比较两个方法的性能如何,然后作出综合的分析.这避开了理论上难于解决的困难问题,有极大的实用意义.

这种情况的出现,也给统计学的发展提出了问题.从大处说,它难免使人觉得传统的统计学理论的作用降低了.因为,既然计算机可以解决一些以往需要用理论解决的问题,那么,发展理论是否就变得不那么迫切了.我们对这个问题的看法是否定的.不错,在有些情况下,数据本身就不好看成是从一定的统计总体中抽出的(这主要

因为它们原则上不能在同样的条件下重复),但用计算机去处理,寻找数据中所包含的规律性和提取其中的信息,仍是有意义的工作.这一点不能看成是计算机的应用取代了数理统计学的理论.因为,按照对统计学意义的现行理解,这种数据的分析问题本来就不属于统计学的范围.而在这个范围内,虽则在某些技术问题上,计算机的使用确能解决以往需要用复杂理论解决的问题.但在涉及一个统计方法的全面性质,以及几种统计方法优良性的比较等问题上,计算机并不能代替理论的作用.因为只能选择有限组参数值去进行模拟,这种模拟的结果可以指示结论可能的性质,但不能据此下定论.实际上,可以说情况正好相反:计算机的使用,给统计学的理论提出了一些新的研究课题.举例言之,在数据分析中,发展了一种方法,叫"投影追踪法",或简称"PP 方法".这种方法的精神,是通过把高维数据向低维空间投影,寻找在某种意义下最好的"投影方向",以便把复杂的高维问题转化为较易处理的低维问题.目前,通过计算机模拟,已证实了在一定情况下,这种方法比之传统方法确有其优越性.但随着这方法的发展,也提出了一些重要的理论问题,只有解决了这个问题,这种方法才可能站稳脚根,并为人们所真正接受.可以说,这个状况与计算机的使用对计算数学的影响相似:计算机的使用淘汰了一些过时的计算方法,但也给计算方法的理论研究提出了不少的新课题.

第四是瓦尔德(A. Wald)的统计决策理论的提出,以及贝叶斯(T. Bayes)统计学派的进展.瓦尔德(1902～

1950)是原籍罗马尼亚的美国统计学家，是本世纪中对统计学面貌的改观起了重大影响的少数几个统计学家之一．1950年，他发表了题为《统计决策函数》的著作，正式提出了这个理论．瓦尔德本来的想法，是要把统计学的各分支都统一在“人与大自然的博奕”这个模式下，以便作出统一的处理．例如，参数估计和假设检验，看起来是差别很大的分支，但在瓦尔德的理论中，形式地统一起来了．他这个理论引起统计学界很大的兴趣．不过，往后的发展表明，瓦尔德最初的设想并未取得很大的成功，但却有着两方面的重要影响：一是瓦尔德把统计推断的后果与经济上的得失联系起来，这使统计方法更便于直接用到经济性的决策的领域；二是瓦尔德理论中所引进的许多概念和问题的新提法，丰富了以往的统计理论．例如参数的点估计理论，在战后时期受到瓦尔德理论很大的影响，以致其面貌有了很大的改变．其他统计分支，也程度不同地受到他的理论的影响．因此，把瓦尔德列为对近代统计学作出重大贡献的学者之一，是当之无愧的．

贝叶斯统计学派的基本思想，源出于英国学者贝叶斯（1702～1761）的一项工作，发表于他去世后的1763年．后世的学者把它发展为一整套关于统计推断的系统理论．信奉这种理论的统计学者，就组成了贝叶斯学派．这个理论在两个方面与传统理论（即基于概率的频率解释的那种理论）有根本的区别：一是否定概率的频率解释，这涉及与此有关的大量统计概念，而提倡给概率以“主观上的相信程度”这样的解释；二是“先验分布”的使

用，先验分布被理解为在抽样前对推断对象的知识的概括. 按照贝叶斯学派的观点，样本的作用，在于且仅在于对先验分布作修改，而过渡到“后验分布”——其中综合了先验分布中的信息与样本中包含的信息. 在此，因篇幅关系，不能对此作详细解释，并将其与传统理论作仔细的比较；只指出这种方法在应用上方便，在一些情况下，其意义也显得更自然且易于了解. 所以，直到本世纪四十年代，这个学派一直未得到重大进展，但近几十年来情况有了很大的改变，其信奉者愈来愈多，其中包括一些有影响的学者. 这两个学派之间的争论，是战后时期统计学的一个重要特点. 在这种争论中，提出了不少问题促使人们进行研究，其中有的是很根本性的，例如，对统计推断和主观概率这种基本概念的深入研究，促进了统计学的发展，有人且认为将对未来的统计学的面貌起重要影响. 贝叶斯学派与瓦尔德统计决策理论的联系在于：这二者的结合，产生了“贝叶斯决策理论”，它构成统计决策理论在实际应用上的主要内容.

以上是关于统计学现状及存在的问题的一个很粗略的介绍. 应当指出，这一叙述在一定程度上只是作者个人的主观看法，不一定都很妥当或确切.

最后谈谈我国统计学发展的简单情况. 这里，“统计学”一词是在本书前述意义下去理解的.

我国现代统计学的研究起步较晚. 本世纪三十年代，有许宝騄等人去当时统计学最发达的国家——英国，随费歇耳等著名统计学家学习和进行研究工作. 其中作出

了杰出贡献，并取得国际影响的，有许宝騄教授. 许教授在三、四十年代发表了一系列重要论文，涉及统计学理论的一些领域，其中尤以在多元分析和线性模型的统计推断方面的工作最为突出，有的且是奠基性的.许教授的一些工作，直到现在还常被引用，可以说已成为本学科方面的经典著作. 许教授在培养统计学人才方面也作出了重大贡献. 他曾执教于美国，其学生中，有的后来成为美国统计学界的权威和前辈.在国内，他曾执教于西南联合大学和北京大学——解放后至他去世的1970年为止，他一直在北京大学，除了继续进行统计学理论的研究工作外，更重要的是他培养了一批学生，其中不少现已成为我国统计学界的骨干.总之，许教授对世界和我国统计学的发展，都作出了巨大的贡献.

在我国，统计学一直被看成数学的一个分支.但统计学中所用的数学工具比较初浅古典，与近代数学发展的主流相去甚远，纯数学方面的学者多不愿问津，加上我国工农业生产、科技方面的落后，以及其他种种原因，统计学的发展缺乏应用方面需求的推动.这种种情况，使得我国统计学的发展，与国际先进水平相比落后很多.到解放时的情况是：国内从事这方面研究工作的，只有屈指可数的几个人；高等学校除个别例外，都不能开出这方面的课程；应用和出版方面的情况就更差.

解放后，在中国科学院数学研究所中，先后建立了概率统计的研究组和室(1980年数学所改组为三个所后，统计学方面的研究人员分别安排在系统科学研究所和应用

数学研究所内).1956年,在制定科学规划时,对这学科给予较大的关注.中国科学院内,这方面的研究队伍有所扩大.个别条件较好的高等学校,也加强了本学科的人才培养工作. 自1955年到1966年,一批中、青年学者在研究工作中取得了成绩. 其中如王寿仁、张里千、成平、张尧庭、刘璋温等人,在非参数统计、参数估计和试验设计方面,作出了达到或接近国际先进水平的工作.自1958年以来,统计工作者开始较大量地将统计方法应用到国民经济和科技领域中,并取得了初步的成绩和经验. 以上这些,为我国统计学的大发展打下了初步的基础. 在十年动乱期间,理论方面的研究工作和培养人才的工作基本上陷于停顿,我国与国外先进水平的差距拉大了;但在应用方面,这期间还是有一些进展的,例如,把统计方法用于工农业(主要是工业)得到更大的推广,其他如在地质、医药、气象、地震和水文预报方面,都开展了一些工作,并取得效果.

1977年以来,我国统计学的发展进入了一个新时期.统计学研究的队伍扩大了,在一些分支中作出了高质量的研究工作. 在培养人才方面,目前已有五所高等学校(北京大学、复旦大学、南开大学、武汉大学、华东师范大学建立了这方面的系,保证了按本学科的特点去培养人才. 到目前为止,概率统计方面已获得博士学位的,近十名;已获得硕士学位的,百余人. 教师队伍的量和质,也有很大的提高. 应用方面,无论在广度和深度方面,都有进展,作出了一些获得各级奖励的工作. 出版方面,

我国近年来出版了近百种专著和教科书，专业杂志上登载的统计学论文也日渐增多，并从 1985 年 9 月起刊行了我国第一本概率统计方面的专业学术杂志——《应用概率统计》. 从这个势头来看，虽则目前我们与国际先进水平尚有较大的差距，但只要有正确的政策，并坚持不懈的努力，随着我国经济的振兴，统计学这门有重大应用价值的学科必能以较快的速度发展，迎头赶上世界先进水平. 让所有有志于这个学科的人一起努力吧.